SUPREME COURT FOR THE STATE OF NEW YORK
COUNTY OF ALBANY

JOHN KEONI WRIGHT; GINET BORRERO; TAUANA GOINS; NINA DOSTER; CARLA WILLIAMS; MONA PRADIA; ANGELES BARRAGAN,

Plaintiffs,

- against -

STATE OF NEW YORK; REGENTS OF THE UNIVERSITY OF THE STATE OF NEW YORK; MERRYL H. TISCH, CHANCELLOR OF THE BOARD OF REGENTS; JOHN B. KING, COMMISSIONER OF EDUCATION AND PRESIDENT OF THE UNIVERSITY OF THE STATE OF NEW YORK,

Defendants.

Index No.:
Date purchased: July 28, 2014

Summons

Plaintiffs designates Albany County as the place of trial.

The basis for venue is CPLR §§ 503(a), 505

Defendants reside at

County of Albany

To the State of New York:

You are hereby summoned to answer the complaint in this action and to serve a copy of your answer, or, if the complaint is not served with this summons, to serve a notice of appearance, on the Plaintiff's Attorneys within 20 days after the service of this summons, exclusive of the day of service (or within 30 days after the service is complete if this summons is not personally delivered to you within the State of New York); and in case of your failure to appear or answer, judgment will be taken against you by default for the relief demanded in the complaint.

Dated: July 28, 2014

Attorneys for the Plaintiffs

KIRKLAND & ELLIS LLP

Jay P. Lefkowitz
lefkowitz@kirkland.com
Devora W. Allon
devora.allon@kirkland.com
Danielle R. Sassoon
danielle.sassoon@kirkland.com
Sarah M. Sternlieb
sarah.sternlieb@kirkland.com
601 Lexington Avenue
New York, New York 10022-4611
Telephone: (212) 446-4800
Facsimile: (212) 446-4900

Defendant's address:

Office of the Attorney General
Justice Building, Second Floor
Empire State Plaza
Albany, NY 12224

SUPREME COURT FOR THE STATE OF NEW YORK
COUNTY OF ALBANY

JOHN KEONI WRIGHT; GINET BORRERO; TAUANA GOINS; NINA DOSTER; CARLA WILLIAMS; MONA PRADIA; ANGELES BARRAGAN, Plaintiffs, - against - STATE OF NEW YORK; REGENTS OF THE UNIVERSITY OF THE STATE OF NEW YORK; MERRYL H. TISCH, CHANCELLOR OF THE BOARD OF REGENTS; JOHN B. KING, COMMISSIONER OF EDUCATION AND PRESIDENT OF THE UNIVERSITY OF THE STATE OF NEW YORK, Defendants.	Index No.: Date purchased: July 28, 2014 **Summons** Plaintiffs designates Albany County as the place of trial. The basis for venue is CPLR §§ 503(a), 505 Defendants reside at County of Albany

To the Regents of the University of the State of New York:

You are hereby summoned to answer the complaint in this action and to serve a copy of your answer, or, if the complaint is not served with this summons, to serve a notice of appearance, on the Plaintiff's Attorneys within 20 days after the service of this summons, exclusive of the day of service (or within 30 days after the service is complete if this summons is not personally delivered to you within the State of New York); and in case of your failure to appear or answer, judgment will be taken against you by default for the relief demanded in the complaint.

Dated: July 28, 2014

Attorneys for the Plaintiffs

KIRKLAND & ELLIS LLP

Jay P. Lefkowitz
lefkowitz@kirkland.com
Devora W. Allon
devora.allon@kirkland.com
Danielle R. Sassoon
danielle.sassoon@kirkland.com
Sarah M. Sternlieb
sarah.sternlieb@kirkland.com
601 Lexington Avenue
New York, New York 10022-4611
Telephone: (212) 446-4800
Facsimile: (212) 446-4900

Defendant's address:

89 Washington Avenue
Board of Regents, Room 110 EB
Albany, New York 12234

SUPREME COURT FOR THE STATE OF NEW YORK
COUNTY OF ALBANY

JOHN KEONI WRIGHT; GINET BORRERO; TAUANA GOINS; NINA DOSTER; CARLA WILLIAMS; MONA PRADIA; ANGELES BARRAGAN, Plaintiffs, - against - STATE OF NEW YORK; REGENTS OF THE UNIVERSITY OF THE STATE OF NEW YORK; MERRYL H. TISCH, CHANCELLOR OF THE BOARD OF REGENTS; JOHN B. KING, COMMISSIONER OF EDUCATION AND PRESIDENT OF THE UNIVERSITY OF THE STATE OF NEW YORK, Defendants.	Index No.: Date purchased: July 28, 2014 **Summons** Plaintiffs designates Albany County as the place of trial. The basis for venue is CPLR §§ 503(a), 505 Defendants reside at County of Albany

To Merryl H. Tisch, Chancellor of the Board of Regents:

You are hereby summoned to answer the complaint in this action and to serve a copy of your answer, or, if the complaint is not served with this summons, to serve a notice of appearance, on the Plaintiff's Attorneys within 20 days after the service of this summons, exclusive of the day of service (or within 30 days after the service is complete if this summons is not personally delivered to you within the State of New York); and in case of your failure to appear or answer, judgment will be taken against you by default for the relief demanded in the complaint.

Dated: July 28, 2014

Attorneys for the Plaintiffs

KIRKLAND & ELLIS LLP

Jay P. Lefkowitz
lefkowitz@kirkland.com
Devora W. Allon
devora.allon@kirkland.com
Danielle R. Sassoon
danielle.sassoon@kirkland.com
Sarah M. Sternlieb
sarah.sternlieb@kirkland.com
601 Lexington Avenue
New York, New York 10022-4611
Telephone: (212) 446-4800
Facsimile: (212) 446-4900

Defendant's address:

89 Washington Avenue
Albany, New York 12234

SUPREME COURT FOR THE STATE OF NEW YORK
COUNTY OF ALBANY

JOHN KEONI WRIGHT; GINET BORRERO; TAUANA GOINS; NINA DOSTER; CARLA WILLIAMS; MONA PRADIA; ANGELES BARRAGAN, Plaintiffs, - against - STATE OF NEW YORK; REGENTS OF THE UNIVERSITY OF THE STATE OF NEW YORK; MERRYL H. TISCH, CHANCELLOR OF THE BOARD OF REGENTS; JOHN B. KING, COMMISSIONER OF EDUCATION AND PRESIDENT OF THE UNIVERSITY OF THE STATE OF NEW YORK, Defendants.	Index No.: Date purchased: July 28, 2014 **Summons** Plaintiffs designates Albany County as the place of trial. The basis for venue is CPLR §§ 503(a), 505 Defendants reside at County of Albany

To John B. King, Jr., Commissioner of Education and President of the University of the State of New York:

You are hereby summoned to answer the complaint in this action and to serve a copy of your answer, or, if the complaint is not served with this summons, to serve a notice of appearance, on the Plaintiff's Attorneys within 20 days after the service of this summons, exclusive of the day of service (or within 30 days after the service is complete if this summons is not personally delivered to you within the State of New York); and in case of your failure to appear or answer, judgment will be taken against you by default for the relief demanded in the complaint.

Dated: July 28, 2014

Attorneys for the Plaintiffs

KIRKLAND & ELLIS LLP

Jay P. Lefkowitz
lefkowitz@kirkland.com
Devora W. Allon
devora.allon@kirkland.com
Danielle R. Sassoon
danielle.sassoon@kirkland.com
Sarah M. Sternlieb
sarah.sternlieb@kirkland.com
601 Lexington Avenue
New York, New York 10022-4611
Telephone: (212) 446-4800
Facsimile: (212) 446-4900

Defendant's address:

89 Washington Avenue
Albany, New York 12234

SUPREME COURT OF THE STATE OF NEW YORK
COUNTY OF ALBANY

JOHN KEONI WRIGHT; GINET BORRERO; TAUANA GOINS; NINA DOSTER; CARLA WILLIAMS; MONA PRADIA; ANGELES BARRAGAN, Plaintiffs, - against - STATE OF NEW YORK; REGENTS OF THE UNIVERSITY OF THE STATE OF NEW YORK; MERRYL H. TISCH, CHANCELLOR OF THE BOARD OF REGENTS; JOHN B. KING, COMMISSIONER OF EDUCATION AND PRESIDENT OF THE UNIVERSITY OF THE STATE OF NEW YORK, Defendants.	Index No. I.A.S. Part: Justice: **COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF**

PRELIMINARY STATEMENT

1. New York's Constitution guarantees all children in the State a sound basic education. Yet in any given school year, New York schoolchildren are at risk of being assigned to an ineffective teacher.

2. A child's teacher is the single most influential school-based variable in the adequacy of the child's education, and a teacher's quality is a critical determinant of a student's educational success. For the all-too-many New York children taught by an ineffective teacher, the damage to their educational advancement is significant and long-lasting.

3. The status quo in New York's education system is neither tolerable nor unavoidable. It is the product of outdated laws that protect ineffective teachers well above what due process requires and at the direct expense of their students' constitutional rights. These laws hamstring school administrators from making employment decisions based on student need and obstruct them from restoring the quality of the New York public education system. Cumulatively, these laws make it nearly impossible to dismiss and discipline teachers with a proven track record of ineffectiveness or misconduct. Plaintiffs, and other New York State schoolchildren, are the primary victims of this failing system.

4. Plaintiff John Keoni Wright's twin daughters, Kaylah and Kyler are New York public school students whose divergent experiences at school exemplify the direct effects that a teacher's quality has on a child's education. Kaylah and Kyler share nearly everything in common, including their birth date and home life. But one variable separates their life experiences and futures: last year, Kyler was assigned to an ineffective teacher.

5. The effects are apparent. In one year alone, the difference in the twins' teachers caused measurable differences in their educational progress. Kaylah excelled with the benefit of an effective teacher, while Kyler fell behind and is still struggling to catch up with her twin. In terms of reading skills alone, Kaylah and Kyler are now reading several levels apart. The gulf between Kaylah's and Kyler's learning illustrates what is a matter of common sense. An ineffective teacher can leave a student ill-equipped to advance, or even to stay apace of those alike in all respects except the quality of their teacher.

6. This suit challenges the constitutionality, in whole or in part, of Education Laws §§ 2509, 2510, 2573, 2585, 2588, 2590, 3012, 3012-c, 3020, and 3020(a) (the "Challenged Statutes"). The Challenged Statutes confer permanent employment, prevent the removal of ineffective teachers from the classroom, and mandate that layoffs be based on seniority alone, rather than effectiveness. These Statutes prevent students like Kyler Wright and the other plaintiffs from obtaining the sound basic education guaranteed under Article XI, § 1 of the New York Constitution (the "Education Article").

7. This suit seeks to strike down the legal impediments that prevent New York's schools from providing a sound basic education to all of their students, as guaranteed by the New York Constitution. Plaintiffs seek a declaration that the Challenged Statutes violate the constitutional rights of New York schoolchildren and a permanent injunction to prevent their future enforcement.

JURISDICTION AND VENUE

8. Venue is proper in the County of Albany pursuant to the Civil Practice Law and Rules 503(a) and 505(a) because the Defendants' principal offices are located in the County of Albany.

9. The Supreme Court has jurisdiction to hear this case and grant declaratory judgment and appropriate injunctive relief pursuant to Civil Practice Law and Rules 3001 and 3017(b).

PARTIES

Plaintiffs

10. Plaintiff John Keoni Wright sues on his own behalf and on behalf of his minor children, Kaylah and Kyler Wright, students who attend P.S. 158, a Brooklyn school in the New York City School District.

11. Plaintiff Ginet Borrero sues on her own behalf and on behalf of her minor child, Raymond Diaz, Jr., a student who attends I.S. 171, a Brooklyn school in the New York City School District.

12. Plaintiff Tauana Goins sues on her own behalf and on behalf of her minor child, Tanai Goins, a student who attends P.S. 106, a Queens school in the New York City School District.

13. Plaintiff Nina Doster sues on her own behalf and on behalf of her minor children, Patience and King McFarlane, students who attend P.S. 140, a Queens school in the New York City School District.

14. Plaintiff Carla Williams sues on her own behalf and on behalf of her minor child, Jada Williams, a student who previously attended Nathaniel Rochester Community School No. 3 in the Rochester City School District and now attends World of Inquiry School No. 58 in the Rochester City School District.

15. Plaintiff Mona Pradia sues on her own behalf and on behalf of her minor child, Adia-Jendayi Pradia, a student who previously attended Audubon School No. 333 in the Rochester City School District and now attends Norman Howard School, paid for by the Rochester City School District.

16. Plaintiff Angeles Barragan sues on her own behalf and on behalf of her minor child, Natalie Mendoza, a student who attends P.S. 94, Kings College Elementary School, a Bronx school in the New York City School District.

Defendants

17. Defendant the State of New York (the "State") is responsible for the educational system in New York.

18. Defendant Regents of the University of the State of New York ("Board of Regents") is an executive department of the State of New York. The Board of Regents is empowered by the New York Legislature to determine educational policy and promulgate rules to effectuate New York State education law and policies.

19. Defendant Merryl H. Tisch is the Chancellor of the Board of Regents. As Chancellor, Ms. Tisch is the head of the Board of Regents and presides over Regents meetings

and appoints its committees. N.Y. Educ. L. § 203; 8 NYCCR 3.1(a). She is sued in her official capacity.

20. Defendant John B. King, Jr. is the Commissioner of Education and President of the University of the State of New York. As Commissioner, Mr. King has the obligation and authority to supervise and monitor all public schools and to assure that educational services are being provided in New York as required by law and regulation. N.Y. Educ. L. §§ 302-03, 305(2), 308. He is sued in his official capacity.

21. Collectively, the defendants are legally responsible for the operation of the New York State educational system and are required to ensure that its operation complies with relevant state and federal constitutional requirements.

BACKGROUND

22. The Education Article provides that "[t]he legislature shall provide for the maintenance and support of a system of free common schools, wherein all the children of this state may be educated." N.Y. Const. Art. XI, § 1. Article XI guarantees all students in New York a sound basic education. A sound basic education is the key to a promising future, preparing children to realize their potential, be productive citizens, and contribute to society.

23. The State fails to meet its constitutional obligation when it provides deficient inputs to adequately educate its students. Students are entitled to adequate teaching by effective personnel because teachers are the core "input" of a sound basic education.

24. The New York Legislature enacted the Challenged Statutes. Through enforcement by the Defendants, the Challenged Statutes confer permanent employment, prevent

the removal of ineffective teachers, and result in layoffs of effective teachers in favor of less-effective, more senior teachers. Under the existing tenure laws, teachers are granted essentially permanent employment before their effectiveness can be determined. The current dismissal and disciplinary laws for tenured teachers make it nearly impossible to remove ineffective teachers from the classroom once they are prematurely tenured.

25. Because of the Challenged Statutes, New York schoolchildren are taught by ineffective teachers who otherwise would not remain in the classroom. These laws prevent school administrators from dismissing and disciplining teachers who do not meet the most basic standards of adequacy and effectiveness, and from making employment decisions driven by their students' constitutional right to a sound basic education.

26. The State's promotion and retention of ineffective teachers, through its promulgation and enforcement of the Challenged Statutes, violates the New York Constitution.

I. TEACHER EFFECTIVENESS IS A NECESSARY INPUT TO A SOUND BASIC EDUCATION.

27. Effective teachers are the most important factor in student performance. Recent studies have confirmed what the Court of Appeals recognized over ten years ago: teachers "are the first and surely the most important input" in creating an adequate education. *Campaign for Fiscal Equality, Inc. v. State (CFE II)*, 100 N.Y.2d 893, 909 (2003).

28. The key determinant of educational effectiveness is teacher quality. (*See, e.g.*, Ex. 1, Chetty et al., Nat'l Bureau of Econ. Research, *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood* (2011).)

29. In the short-term, effective teachers provide tangible educational results in the form of higher test scores and higher graduation rates. (Ex. 2, Bill & Melinda Gates Found., *Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study* (2013); Ex. 3, Eric A. Hanushek, *Valuing Teachers: How Much Is a Good Teacher Worth?*, Education Next, Summer 2011, at 42.)

30. In the long-term, students taught by effective teachers are given a strong foundation from which to advance and achieve. These students are less likely to become teenage parents and more likely to progress in their education, attending college and matriculating at colleges of higher quality. They are more likely to earn more money throughout their lives, live in neighborhoods of higher socioeconomic status, and save more money for retirement. (*See* Ex. 1, Chetty et al., *supra.*)

31. Teacher quality affects student success more than any other in-school factor. High-quality instruction from effective teachers helps students overcome the traditional barriers demographics impose, (*see* Ex. 4, Steven G. Rivkin et al., *Teachers, Schools, and Academic Achievement*, 73 Econometrica 417, 419 (2005)), and may have the greatest positive effect on low-performing students and minorities, (*see* Ex. 5, Daniel Aaronson et al., *Teachers and Student Achievement in the Chicago Public High Schools*, 25 J. Lab. Econ. 95, 126-128 (2007)).

32. If schools were able to replace the least effective teachers, it would add enormous value to the future earnings of students and the U.S. economy as a whole. (Ex. 3, Hanushek, *supra*, at 43-44.)

33. In light of the substantial and enduring effect that teachers have on their students' achievement, the ability to remove ineffective teachers employed by the New York public school

system would improve the lives and better the futures of the students who would otherwise be assigned to those teachers. Yet the Challenged Statutes deprive New York students of a sound basic education, providing no true means for administrators to remove teachers with a track record of ineffectiveness, and causing too many students to remain in the classroom with ineffective teachers.

II. THE TEACHER TENURE STATUTES CONFER PERMANENT EMPLOYMENT ON INEFFECTIVE TEACHERS.

34. Sections 2509, 2573, 3012 and 3012-c (the "Permanent Employment Statutes"), alone and in conjunction with the other statutes at issue, ensure that ineffective teachers unable to provide students with a sound basic education are granted virtually permanent employment in the New York public school system and near-total immunity from termination.

35. New York Education Law § 3012(2)[1] provides that "at the expiration of the probationary term of a person appointed for such term, subject to the conditions of this section, the superintendent of schools shall make a written report to the board of education or the trustees of a common school district recommending for appointment on tenure those persons who have been found competent, efficient and satisfactory, consistent with any applicable rules of the board of regents adopted pursuant to section 3012(b) or this article."

36. Tenure confers extraordinary benefits and protections, but it is out of the ordinary for a teacher to be denied tenure. The default is to grant teachers tenure and the process is a formality, rather than an appraisal of teacher performance. (*See* Ex. 6, Ann Duffett et al., Educ.

---

[1] Section 3012 applies to certain school districts, including common school districts and/or school districts employing fewer than eight teachers, other than city school districts. Section 2509 applies the same law to school districts of cities with less than 125,000 inhabitants. Section 2573 applies the same law to school districts of cities with 125,000 inhabitants or more.

Sector, *Waiting to Be Won Over: Teachers Speak on the Profession, Unions, and Reform* 3 (2008).)

37. In 2007, 97 % of tenure-eligible New York City teachers received tenure. Even with recent reforms meant to strengthen the evaluation system, few teachers are denied tenure. In 2011 and 2012, while some teachers had their probationary periods extended, only 3 % of tenure-eligible teachers were denied tenure outright. (*See* Ex. 7, Susanna Loeb et al., *Performance Screens for School Improvement: The Case of Teacher Tenure Reform in New York City* (2014).) These numbers indicate that most ineffective teachers are not denied tenure.

38. New York school districts typically grant tenure to new teachers after a probationary period of three years, and after only two years of performance review. The statute's prescribed methods for evaluating effectiveness before granting tenure are deficient and three years is inadequate to assess whether a teacher has earned the lifelong benefits of tenure.

39. Pursuant to New York Education Law § 3012-c(1), New York State implemented the Annual Professional Performance Review (the "APPR") to evaluate teachers and principals. A teacher's review is meant to be a significant factor in employment decisions, including tenure, retention, and termination. N.Y. Educ. Law. § 3012-c(1).

40. Under the APPR, teachers receive a numerical score every year that is transposed into one of four ratings: "Highly Effective," "Effective," Developing," or Ineffective." Each school district negotiates the specific terms of their APPR plans, which must comply with § 3012-c. State-developed measures of student growth, such as test results, must form twenty percent of a teacher's rating. Another twenty percent must be based on locally selected measures of student achievement. Locally determined evaluation methods, such as classroom observations

by administrative staff, form the remaining sixty percent. Rather than impose a uniform definition of what constitutes conduct unworthy of tenure, the Permanent Employment Statutes have invited variable and superficial definitions of ineffective teaching that do not ensure tenure is awarded only to effective teachers.

41. The APPR does not adequately identify teachers who are truly "Developing" or "Ineffective." For example, teachers are not rated ineffective even when their students consistently fail state exams. In 2012, only 1 % of teachers were rated "Ineffective."[2] At the same time, 91.5 % of New York teachers were rated "Highly Effective" or "Effective," even though only 31 % of students taking the English Language Arts and Math standardized tests met the standard for proficiency. (Ex. 8, Cathy Woodruff, *Why Are Most Teachers Rated Effective When Most Students Test Below Standards?*, N.Y. St. Sch. Bds. Ass'n, (Dec. 16, 2013), http://www.nyssba.org/news/2013/12/12/on-board-online-december-16-2013/why-are-most-teachers-rated-effective-when-most-students-test-below-standards/.)

42. Similarly, of the New York City teachers eligible for tenure from 2010-11 to 2012-2013, only 2.3 % received a final rating of "Ineffective" (302 teachers), even though 8 % of the teachers had low attendance (more than twenty absences over prior two years) and 12 % of teachers had low value added. (*See* Ex. 7, Loeb et al., *supra*.) These discrepancies indicate that the APPR ratings operate as a rubber stamp for tenure and are not a meaningful check within the tenure process.

---

[2] The data excludes New York City teachers because the city and teachers' union were unable to agree on a plan for the teacher evaluation system. (Ex. 9, Geoff Decker, *Few Teachers Across New York State Earned Low Ratings Last Year*, Chalkbeat, (Oct. 22, 2013), http://ny.chalkbeat.org/2013/10/22/few-teachers-across-new-york-state-earned-low-ratings-last-year/#.U3oacPldXgU.) On information and belief, the New York City data would be similar to the overall New York State data.

43. The APPR's deficient and superficial means of assessing teacher effectiveness is the most highly predictive measure of whether a teacher will be awarded tenure. (*See id.*)

44. The few teachers receiving an "Ineffective" or "Developing" rating are not the only ineffective teachers in the New York public school system. It is less likely that so few teachers are ineffective than that the ratings of many ineffective teachers are inflated and the ineffective performance by teachers is roundly ignored. The ratings do not identify pedagogically incompetent teachers, including teachers unable to control their classroom, who fail to provide instruction, prepare lesson plans, or distribute homework, and teachers indifferent to their students' educational advancement.

45. Of the miniscule percentage of ineffective teachers actually rated as such, not all are denied tenure. Between 2010 and 2013, close to 1 % were approved for tenure and 18.2 % had their probationary periods extended. (*See id.*) In addition, teachers have the right to appeal an Ineffective rating[3] and tenure cannot be denied to a probationary teacher while an APPR appeal about the teacher's performance is pending. N.Y. Educ. Law § 3012-c(5). Moreover, administrators renew probationary teachers in their final probationary year despite any performance concerns. (Ex. 11, Communities for Teaching Excellence, *Earned, Not Given: Transforming Teacher Tenure* 3 (2012).)

46. A teacher's long-term effectiveness cannot be determined with any degree of confidence during the first two or three years of teaching. Most studies indicate that teacher effectiveness is typically established by the fourth year of teaching. (*Id.* at 5.) After that,

---

[3] Most districts also allow tenured, as well as non-tenured, teachers to appeal a Developing rating. (*See* Ex. 10, Alexander Colvin et al., Scheinman Inst. on Conflict Resolution, *APPR Teacher Appeals Process Report* (2014).)

effective teachers tend to remain relatively effective, and ineffective teachers remain relatively ineffective. Deciding tenure after a three-year probationary period confers permanent employment on many teachers who will be ineffective for the rest of their teaching career.

47. The statute's notification requirements make it effectively impossible to consider a teacher's third-year APPR before a tenure determination is made, even if a teacher is found to be ineffective in the third year of his or her probationary period. Section 3012 requires the superintendent of school to notify in writing "each person who is not to be recommended" for tenure of that decision no later than sixty days before the expiration of his or her probationary period. N.Y. Educ. Law § 3012(2). Typically, however, a teacher's probationary term ends before the third-year APPR is reported, at the end of the school year. (*See* Ex. 12, Warren H. Richmond III, *Evaluation Law Could Limit Ability to Terminate Probationary Teachers*, N.Y.L.J., May 16, 2013, at 2.) The final APPR rating may not be provided until September 1 of the following school year. N.Y. Educ. Law § 3012-c(2)(c)(2). A tenure determination, therefore, may be made on the basis of only two years of APPR reviews, and without regard to an ineffectiveness determination in the third year.

48. Once a teacher receives tenure, he or she is guaranteed continued employment except in limited enumerated circumstances and only after a disciplinary hearing pursuant to section 3020(a).

III. THE DISCIPLINARY STATUTES KEEP INEFFECTIVE, TENURED TEACHERS IN THE SCHOOL SYSTEM.

49. Once a teacher receives tenure, he or she cannot be removed except for just cause, and in accordance with the disciplinary process prescribed by § 3020-a. N.Y. Educ. Law § 3020(1) (§ 3020-a and § 3020 hereinafter collectively referred to as the "Disciplinary

Statutes"). The following causes may constitute reason to remove or discipline a teacher: insubordination, immoral character or conduct unbecoming of a teacher, inefficiency, incompetency, physical or mental disability, or neglect of duty, or a failure to maintain required certification. N.Y. Educ. Law § 3012(2).

50. As applied, the Disciplinary Statutes result in the retention of ineffective teachers. The Disciplinary Statutes impose dozens of hurdles to dismiss or discipline an ineffective teacher, including investigations, hearings, improvement plans, arbitration processes, and administrative appeals. On top of these procedural obstacles, the standard for proving just cause to terminate a teacher is nigh impossible to satisfy. The statutorily mandated hearings are "consuming and expensive hurdles that make the dismissal of chronically ineffective, tenured teachers almost impossible." (Ex. 11, Communities for Teaching Excellence, *supra*, at 5.)

51. The Disciplinary Statutes make it prohibitively expensive, time-consuming, and effectively impossible to dismiss an ineffective teacher who has already received tenure. Because of the difficulty, cost, and length of time associated with removal, the number of ineffective teachers who remain employed is far higher than the number of those disciplined or terminated.

52. Disciplinary proceedings are rarely initiated. It is well known that "because of the cumbersome, lengthy, and costly due process protections [tenure] affords, many school districts rarely attempt to fire teachers--in effect granting them permanent employment." (*Id.* at 2.)

53. As an initial matter, administrators are deterred from giving an Ineffective rating. On information and belief, principals and other administrators may be inclined to rate teachers artificially high because of the lengthy appeals process for an ineffectiveness rating and because

they must partake in the development and execution of a teacher improvement plan ("TIP") for Developing and Ineffective teachers. N.Y. Educ. Law § 3012-c(4). The TIP must be mutually agreed upon by the teacher and principal and must include "needed areas of improvement, a timeline for achieving improvement, the manner in which improvement will be assessed, and, where appropriate, differentiated activities to support a teacher's or principal's improvement in those areas." *Id.*

54. Section 3020-a imposes a three-year limit for bringing charges against a teacher. But before administrators may initiate proceedings to discipline or terminate an ineffective or incompetent teacher, they must meticulously build a trove of evidence that includes extensive observation, detailed documentation, and consultation with the teacher. On information and belief, it may be difficult for school districts to collect enough evidence for a 3020-a hearing within the three-year period. This laborious and complicated process deters administrators from trying to remove ineffective teachers from the classroom. (*See* Ex. 13, John Stossel, *How to Fire an Incompetent Teacher*, Reason (Oct. 2006), http://cloudfront-assets.reason.com/assets/db/12639308918768.pdf.)

55. On information and belief, principals and administrators would be more likely to use the 3020-a process to discipline or dismiss a teacher if it was less time-consuming and more effective. A 2009 survey found that 48 % of districts surveyed considered bringing 3020-a charges at least once, but did not. The districts stated multiple reasons for not filing charges, including that the process was too cumbersome, too expensive, that their case was not strong enough, or that the employee resigned. (*See* Ex. 14, Patricia Gould, *3020-a Process Remains Slow, Costly*, N.Y. St. Sch. Bds. Ass'n (May 11, 2009),

http://www.nyssba.org/index.php?src=news&refno=853&category=On%20Board%20Online%20May%2011%202009.)

56. Once an administrator clears the hurdles to file charges, termination can result only after a 3020-a hearing. Despite statutory time limits, from 2004-2008, 3020-a disciplinary proceedings took an average of 502 days, from the time charges were brought until a final decision. (*See* Ex. 15, 3020-a Teacher Discipline Reform, N.Y. State Sch. Bds. Ass'n, http://www.nyssba.org/index.php?src=gendocs&ref=3020-a%20Teacher%20Discipline%20Reform&category=advocacy_legislation.)[4]

57. Incompetency proceedings, which may include charges such as inability to control a class and failure to prepare required lesson plans, take even longer. From 1995-2006, incompetency proceedings in New York took an average of 830 days, costing $313,000 per teacher. (*Id.*)

58. Two consecutive Ineffective ratings constitute a pattern of ineffective teaching or performance, subjecting a teacher to an expedited § 3020-a hearing. N.Y. Educ. Law § 3020-a(3)(c)(i-a)(A). But few teachers receive two consecutive Ineffective ratings to trigger an expedited process.

59. While charges are pending, ineffective teachers continue to be paid even if they are suspended. Unless a teacher is convicted of certain felony crimes, the teacher "may be suspended pending a hearing on the charges and the final determination thereof" *with pay*. N.Y. Educ. Law § 3020-a(2)(b).

---

[4] The statistics in paragraphs 56-57 exclude New York City, which has an alternate disciplinary process.

60. The Disciplinary Statutes require the following procedure to discipline a teacher:

61. Section 3020(1) incorporates the "alternate disciplinary procedures contained in a collective bargaining agreement." N.Y. Educ. Law § 3020(1). This means that the Statute allows its procedural requirements to be modified by contract. In practice, the collective bargaining agreements make it even more difficult to remove ineffective teachers and add conditions that delay the process even further. For example, in New York City the arbitrator must be jointly selected with the union, which effectively grants the union the power to veto

arbitrators on the list. The refusal to appoint hearing officers contributes to the massive backlog of disciplinary cases in New York City.

62. These proceedings are not only long, they are futile. When administrators do pursue disciplinary action, few 3020-a proceedings result in termination, *even when an arbitrator determines that the teacher is ineffective, incompetent, or has engaged in misconduct.* In a study of New York City 3020-a proceedings from 1997-2007, only twelve teachers were dismissed for incompetent teaching. (Ex. 16, Katharine B. Stevens, *Firing Teachers: Mission Impossible*, N.Y Daily News (Feb. 17, 2014), http://www.nydailynews.com/opinion/firing-teachers-mission-impossible-article-1.1615003.)

63. On information and belief, dismissals are so rare not because there are no incompetent teachers, but because the Permanent Employment and Disciplinary Statutes make it impossible to fire them.

64. Thus, if administrators are ever able to comply with the myriad procedural requirements that precede disciplinary action, they then confront a burden of proof that is nearly insurmountable. In order to terminate a teacher, administrators must not only validate the charges, but also prove that the school has undertaken sufficient remediation efforts, that all remediation efforts have failed, and that they will continue to fail indefinitely. *See, e.g.*, *deSouza v. Dep't of Educ.*, 28 Misc. 3d 1201(A) (N.Y. Sup. 2010).

65. The result of these proceedings is that ineffective teachers return to the classroom, and students are denied the adequate education that is their right.

IV. THE LIFO STATUTES REQUIRE THE STATE TO RETAIN MORE SENIOR TEACHERS AT THE EXPENSE OF MORE EFFECTIVE TEACHERS.

66. When school districts conduct layoffs that reduce the teacher workforce, New York Education Law § 2585 mandates that the last teachers hired be the first teachers fired (the "Last In First Out" or "LIFO" Statute).[5] Under the LIFO Statute, "[w]henever a board of education abolishes a position under this chapter, the services of the teacher having the least seniority in the system within the tenure of the position abolished shall be discontinued." N.Y. Educ. Law § 2585(3).

67. New York is one of only ten states to conduct layoffs on the basis of seniority alone, irrespective of a teacher's performance, effectiveness, or quality. (Ex. 17, *Vergara v. California*, No. BC484642 (Cal. Super. Ct. June 10, 2014).)

68. Under the LIFO Statute, school districts conducting layoffs must fire, junior high-performing teachers. While these teachers are lost to the classroom, senior, low-performing, and more highly-paid teachers continue to provide poor instruction to their students.

69. Seniority is not an accurate predictor of teacher effectiveness. Studies demonstrate that a teacher's effectiveness generally levels off of returns to experience after five to seven years. (Ex. 18, Allison Atteberry et al., *Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness* 4 (CALDER, Working Paper No. 90, 2013).) Yet the LIFO Statute requires that seniority, which has little correlation to a teacher's effectiveness, be the sole factor in layoffs.

---

[5] Section 2585 applies to school districts of cities with 125,000 inhabitants or more, such as Rochester City School District. Section 2510(1)-(2) applies the same law to school districts of cities with less than 125,000 inhabitants. Section 2588 applies to school districts of cities with over 1,000,000 inhabitants, such as New York City.

70. In recent years, various school districts in New York, including the Rochester City School District, have implemented district-wide layoffs due to budgetary constraints. In Rochester, the district laid off 116 teachers in 2010, 400 teachers in 2011, and 56 teachers in 2012. Pursuant to the LIFO Statute, school administrators discontinued the employment of top-performing teachers with lower seniority, and retained low-performing teachers with greater seniority.

71. Under a seniority-based layoff system, school districts must fire more teachers to satisfy budgetary constraints because newer teachers are paid less. The higher the number of layoffs, the greater the detriment suffered by schools and students.

72. Seniority-based layoffs affect children at struggling schools the most, because lower-performing schools generally have a disproportionate number of newly-hired teachers.

73. The LIFO Statute hinders recruitment of talented personnel because newly-hired teachers face a heightened risk of being laid off, regardless of their abilities and performance.

74. Layoffs determined on the basis of teacher effectiveness, rather than seniority alone, would result in a more effective workforce. If New York City had conducted seniority-based layoffs between 2006 and 2009, none of the New York City teachers that received an Unsatisfactory[6] rating during those years would have been laid off. In the absence of the LIFO Statute, school administrators conducting layoffs would consider teacher performance, a higher number of effective teachers would be retained, and fewer children would suffer the loss of an

---

[6] New York changed their rating system in 2010, from rating teachers as 'Satisfactory' or 'Unsatisfactory,' to 'Highly Effective,' 'Effective,' 'Developing,' and 'Ineffective.'

effective teacher. (Ex. 19, Donald Boyd et al., *Teacher Layoffs: An Empirical Illustration of Seniority Versus Measures of Effectiveness*, 6 Educ. Finance & Pol. 439 (2011).)

75. The LIFO Statute, both alone and in conjunction with the other Challenged Statutes, ensures that a number of ineffective teachers unable to provide students with a sound basic education retain employment in the New York school system.

76. Cumulatively, the State's enforcement of the Challenged Statutes forces schools to retain ineffective teachers and violates New York students' right to a sound basic education.

FIRST CAUSE OF ACTION

77. Plaintiffs repeat and re-allege each and every allegation set forth in Paragraphs 1 through 76 as though fully set forth herein at length.

78. The Permanent Employment Statute violates the Education Article of the New York Constitution because it has failed, and continues to fail to provide all children in New York State with a sound basic education.

79. Teacher effectiveness cannot be determined within three years. The teachers who obtain tenure may fail to provide students with an effective education, but are guaranteed lifetime employment and compensation.

SECOND CAUSE OF ACTION

80. Plaintiffs repeat and re-allege each and every allegation set forth in Paragraphs 1 through 76 as though fully set forth herein at length.

81. The Disciplinary Statutes violate the Education Article of the New York Constitution because they fail to provide all children in New York State with a sound basic education by preventing the dismissal of ineffective teachers.

82. Principals are unlikely to take action to attempt to dismiss or discipline an ineffective teacher. Because disciplinary proceedings are time-consuming, costly, and unlikely to result in the removal of teachers, ineffective teachers are kept in the classroom.

THIRD CAUSE OF ACTION

83. Plaintiffs repeat and re-allege each and every allegation set forth in Paragraphs 1 through 76 as though fully set forth herein at length.

84. The LIFO Statute violates the Education Article of the New York Constitution because it has failed, and will continue to fail to provide children throughout the Rochester City School District with a sound basic education.

85. LIFO prohibits administrators from taking teacher quality into account when conducting layoffs so that ineffective, more senior teachers are retained and effective teachers are fired.

PRAYER FOR RELIEF

WHEREFORE, plaintiffs respectfully request that this Court enter a judgment against Defendants as follows:

(i) As to each Count, a declaratory judgment, that the Challenged Statutes violate the New York Constitution in the manner alleged above.

(ii) As to each Count, preliminary and permanent injunctions enjoining Defendants from implementing or enforcing the Challenged Statutes.

(iii) Award plaintiffs all costs and expenses incurred in bringing this action, including reasonable attorney's fees and costs;

(iv) Such other relief available under New York law that may be considered appropriate under the circumstances, and further relief as this Court deems just and proper.

Dated: New York, New York
July 28, 2014

Kirkland & Ellis LLP

By:
Jay P. Lefkowitz

Jay P. Lefkowitz
Devora W. Allon
Danielle R. Sassoon
Sarah M. Sternlieb
KIRKLAND & ELLIS LLP
601 Lexington Ave.
New York, NY 10022
Telephone (212) 446-4800
Facsimile (212) 446-6460

Attorneys for Plaintiffs

EXHIBIT 1
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

NBER WORKING PAPER SERIES

THE LONG-TERM IMPACTS OF TEACHERS: TEACHER VALUE-ADDED AND STUDENT OUTCOMES IN ADULTHOOD

Raj Chetty
John N. Friedman
Jonah E. Rockoff

Working Paper 17699
http://www.nber.org/papers/w17699

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2011

We thank Joseph Altonji, Josh Angrist, David Card, Gary Chamberlain, David Deming, Caroline Hoxby, Guido Imbens, Brian Jacob, Thomas Kane, Lawrence Katz, Adam Looney, Phil Oreopoulos, Jesse Rothstein, Douglas Staiger, Danny Yagan, and seminar participants at the NBER Summer Institute, Stanford, Princeton, Harvard, Univ. of Chicago, Univ. of Pennsylvania, Brookings, Columbia, Univ. of Maryland, Pompeu Fabra, University College London, Univ. of British Columbia, and UC San Diego for helpful discussions and comments. This paper draws upon results from a paper in the IRS Statistics of Income Paper Series entitled "New Evidence on the Long- Term Impacts of Tax Credits on Earnings." Tax microdata were not accessed to write the present paper, as all results using tax data are based on tables contained in the SOI white paper. Peter Ganong, Sarah Griffis, Michal Kolesar, Jessica Laird, and Heather Sarsons provided outstanding research assistance. Financial support from the Lab for Economic Applications and Policy at Harvard and the National Science Foundation is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research. Publicly available portions of the analysis code are posted at: http://obs.rc.fas.harvard.edu/chetty/va_bias_code.zip

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2011 by Raj Chetty, John N. Friedman, and Jonah E. Rockoff. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood
Raj Chetty, John N. Friedman, and Jonah E. Rockoff
NBER Working Paper No. 17699
December 2011, Revised January 2012
JEL No. I2,J24

ABSTRACT

Are teachers' impacts on students' test scores ("value-added") a good measure of their quality? This question has sparked debate largely because of disagreement about (1) whether value-added (VA) provides unbiased estimates of teachers' impacts on student achievement and (2) whether high-VA teachers improve students' long-term outcomes. We address these two issues by analyzing school district data from grades 3-8 for 2.5 million children linked to tax records on parent characteristics and adult outcomes. We find no evidence of bias in VA estimates using previously unobserved parent characteristics and a quasi-experimental research design based on changes in teaching staff. Students assigned to high-VA teachers are more likely to attend college, attend higher- ranked colleges, earn higher salaries, live in higher SES neighborhoods, and save more for retirement. They are also less likely to have children as teenagers. Teachers have large impacts in all grades from 4 to 8. On average, a one standard deviation improvement in teacher VA in a single grade raises earnings by about 1% at age 28. Replacing a teacher whose VA is in the bottom 5% with an average teacher would increase the present value of students' lifetime income by more than $250,000 for the average class- room in our sample. We conclude that good teachers create substantial economic value and that test score impacts are helpful in identifying such teachers.

Raj Chetty
Department of Economics
Harvard University
1805 Cambridge St.
Cambridge, MA 02138
and NBER
chetty@fas.harvard.edu

Jonah E. Rockoff
Columbia University
Graduate School of Business
3022 Broadway #603
New York, NY 10027-6903
and NBER
jonah.rockoff@columbia.edu

John N. Friedman
Harvard Kennedy School
Taubman 356
79 JFK St.
Cambridge, MA 02138
and NBER
john_friedman@harvard.edu

1 Introduction

Many policy makers advocate increasing the quality of teaching, but there is considerable debate about the best way to measure and improve teacher quality. One prominent method is to evaluate teachers based on their impacts on their students' test scores, commonly termed the "value-added" (VA) approach (Hanushek 1971, Murnane 1975, Rockoff 2004, Rivkin, Hanushek, and Kain 2005, Aaronson, Barrow, and Sander 2007, Kane and Staiger 2008). School districts from Washington D.C. to Los Angeles have begun to publicize VA measures and use them to evaluate teachers. Advocates argue that selecting teachers on the basis of their VA can generate substantial gains in achievement (e.g., Gordon, Kane, and Staiger 2006, Hanushek 2009), while critics contend that VA measures are poor proxies for teacher quality and should play little if any role in evaluating teachers (e.g., Baker et al. 2010, Corcoran 2010).

The debate about teacher VA stems primarily from two unanswered questions.[1] First, do the differences in test-score gains across teachers measured by VA capture causal impacts of teachers or are they driven primarily by student sorting? If students are sorted to teachers in ways that are not accounted for when estimating value-added, VA estimates will incorrectly reward or penalize teachers for the mix of students they get. Researchers have reached conflicting conclusions about the degree of bias in VA (e.g. Kane and Staiger 2008, Rothstein 2010) and there is still disagreement about this important issue. Second, do teachers who raise test scores improve their students' outcomes in adulthood or are they simply better at teaching to the test? Recent work has shown that early childhood education has significant long-term impacts (e.g. Heckman et al. 2010a, 2010b, 2010c, Chetty et al. 2011), but no study has identified the long-term impacts of teacher quality as measured by value-added.

We address these two questions using information from two administrative databases. The first is a dataset on test scores and classroom and teacher assignments in grades 3-8 from a large urban school district in the U.S. These data cover more than 2.5 million students and 18 million tests for math and English (reading) spanning 1989-2009. The second is selected data from United States tax records spanning 1996-2010.[2] These data contain information on student outcomes such as earnings, college attendance, and teenage births as well as parent characteristics such as

[1] There are also other important concerns about VA besides the two we focus on in this paper. For instance, as with other measures of labor productivity, the signal in value-added measures may be degraded by behavioral responses if high-stakes incentives are put in place (Barlevy and Neal 2012).

[2] Tax microdata were not directly used to write the present paper, as all results using tax data are drawn from tables contained in a Statistics of Income paper on the long-term impacts of tax policy (Chetty, Friedman, and Rockoff 2011). We describe the details of how the tax data were analyzed here as a reference.

household income, retirement savings, and mother's age at child's birth. We match nearly 90% of the observations in the school district data to the tax data, allowing us to track a large group of individuals from elementary school to early adulthood.

Our analysis has two parts. In the first part, we develop new tests for bias in VA measures. We estimate teacher value-added using standard Empirical Bayes methods, conditioning on predetermined variables from the school district data such as lagged test scores (Kane and Staiger 2008, Kane, Rockoff, and Staiger 2008). Our estimates of VA are consistent with prior work: a 1 standard deviation (SD) improvement in teacher VA raises end-of-grade test scores by approximately 0.1 SD on average. To evaluate whether these VA estimates are biased by sorting on observables, we use parent characteristics from the tax data, which are strong predictors of test scores but are omitted from the VA models. We find that these parent characteristics are uncorrelated with teacher value-added *conditional* on the observables used to fit the VA model from the school district data. In addition, lagged test score gains are essentially uncorrelated with current teacher VA conditional on observables. We conclude that sorting on observable dimensions generates little or no bias in standard VA estimates.

To evaluate sorting on unobservables, we develop a quasi-experimental method of testing for bias in VA estimates that exploits changes in teaching assignments at the school-grade level. For example, suppose a high-VA 4th grade teacher moves from school s to another school in 1995. If VA estimates have predictive content, then students entering grade 4 in school s in 1995 should have lower quality teachers on average and their test score gains should be lower on average than the previous cohort. In practice, we find sharp breaks in test score gains around such teacher arrivals and departures at the school-grade-cohort level. Building on this idea, we assess the degree of bias in VA estimates by testing if observed changes in average test scores across cohorts match predictions based on the changes in the mean value-added of the teaching staff.[3] We find that the predicted impacts closely match observed impacts: the point estimate of the bias in forecasted impacts is 2% and statistically insignificant.[4] Although it rests on stronger identifying assumptions than a randomized experiment, our approach of using variation from teacher turnover

[3] This research design is related to recent studies of teacher turnover (e.g., Rivkin, Hanushek, and Kain 2005, Jackson and Bruegmann 2010, Ronfeldt et al. 2011), but is the first direct test of whether the VA of teachers who enter or exit affects mean test scores across cohorts. We discuss how our approach differs from this earlier work in Section 4.4.

[4] This quasi-experimental test relies on the assumption that teacher departures and arrivals are not correlated at a high frequency with student characteristics. We find no evidence of such correlations based on observables such as lagged test scores or scores in other subjects. This is intuitive, as parents are unlikely to immediately switch their children to a different school simply because a single teacher leaves or arrives.

can be implemented in many datasets and yields much more precise estimates of the degree of bias. Our method requires no data other than school district administrative records, and thus provides a simple technique for school districts and education researchers to validate their own value-added models.[5]

As we discuss in greater detail below, our results reconcile the findings of Kane and Staiger (2008) and Rothstein (2010) on bias in VA estimates. Rothstein finds minimal bias in VA estimates due to selection on observables but warns that selection on unobservables could *potentially* be a problem because students are sorted to classrooms based on lagged gains. Like Rothstein, we find minimal selection on observables. We then directly test for selection on unobservables using an approach analogous to Kane and Staiger (2008), but exploiting quasi-experimental variation in lieu of a randomized experiment. Like Kane and Staiger, we find no evidence of selection on unobservables. We therefore conclude that our value-added measures provide unbiased estimates of teachers' causal impacts on test scores despite the grouping of students on lagged gains documented by Rothstein.[6]

In the second part of the paper, we analyze whether high-VA teachers improve their students' outcomes in adulthood. We structure our analysis using a stylized dynamic model of the education production function in which cumulative teacher inputs over all grades affect earnings, as in Todd and Wolpin (2003). We regress outcomes such as earnings for a given set of students on teacher VA estimated using *other* cohorts to account for correlated errors in scores and earnings, as in Jacob, Lefgren, and Sims (2010). The resulting coefficients capture the "reduced form" impact of being assigned a teacher with higher VA in grade g, which includes both the grade g teacher's direct effect and any indirect benefits of being tracked to better teachers or receiving better educational inputs after grade g.

We first pool all grades to estimate the average reduced-form impact of having a better teacher for a single year from grades 4-8. We find that teacher VA has substantial impacts on a broad range of outcomes. A 1 SD improvement in teacher VA in a single grade raises the probability of college attendance at age 20 by 0.5 percentage points, relative to a sample mean of 36%. Improvements in teacher quality also raise the quality of the colleges that students attend, as measured by the average earnings of previous graduates of that college. Changes in the quality of the teaching

[5] STATA code to implement this technique is available at http://obs.rc.fas.harvard.edu/chetty/va_bias_code.zip

[6] Our findings do not contradict Rothstein's results; in fact, we replicate them in our own data. However, while Rothstein concludes that selection on unobservables could potentially generate significant bias, we find that it is actually negligible based on quasi-experimental tests that provide more definitive estimates of the degree of bias.

staff across cohorts generate impacts on college attendance and quality of a similar magnitude, supporting the view that these estimates reflect the causal impact of teachers.

Students who get higher VA teachers have steeper earnings trajectories, with significantly higher earnings growth rates in their 20s. At age 28, the oldest age at which we have a sufficiently large sample size to estimate earnings impacts, a 1 SD increase in teacher quality in a single grade raises annual earnings by about 1% on average. If this impact on earnings remains constant over the lifecycle, students would gain approximately $25,000 on average in cumulative lifetime income from a 1 SD improvement in teacher VA in a single grade; discounting at a 5% rate yields a present value gain of $4,600 at age 12, the mean age at which the interventions we study occur.

We also find that improvements in teacher quality significantly reduce the probability of having a child while being a teenager, increase the quality of the neighborhood in which the student lives (as measured by the percentage of college graduates in that ZIP code) in adulthood, and raise 401(k) retirement savings rates. The impacts on adult outcomes are all highly statistically significant, with the null of no impact rejected with $p < 0.01$.

Under certain strong assumptions about the nature of the tracking process, the net impacts of teacher VA in grade g can be recovered from the reduced-form coefficients by estimating a set of tracking equations that determine how teacher VA in grade g affects VA in subsequent grades. Using this approach, we find that the net impacts of teacher VA are significant and large throughout grades 4-8, showing that improvements in the quality of education can have large returns well beyond early childhood.[7]

The impacts of teacher VA are slightly larger for females than males. A given increase in test scores due to higher teacher quality is worth more in English than math, but the standard deviation of teacher effects is 50% larger in math than English. The impacts of teacher VA are roughly constant in percentage terms by parents' income. Hence, high income households, whose children have higher earnings on average, should be willing to pay larger absolute amounts for higher teacher VA.

The finding that one's teachers in childhood have long-lasting impacts may be surprising given evidence that teachers' impacts on test scores "fade out" very rapidly in subsequent grades (Rothstein 2010, Carrell and West 2010, Jacob, Lefgren, and Sims 2010). We confirm this rapid fade-out in our data, but find that test score impacts stabilize at about 1/3 the original impact after 3

[7] Because we can only analyze the impacts of teacher quality from grades 4-8, we cannot quantify the returns to education at earlier ages. The returns to better education in pre-school or earlier may be much larger than those estimated here (Heckman 2000).

years, showing that some of the achievement gains persist. Despite the fade-out of impacts on scores, the impacts of better teaching on earnings are similar to what one would predict based on the cross-sectional correlation between earnings and contemporaneous test score gains conditional on observables. This pattern of fade-out and re-emergence echoes the findings of recent studies of early childhood interventions (Heckman et al. 2010c, Deming 2009, Chetty et al. 2011).

To illustrate the magnitude of teachers' impacts, we use our estimates to evaluate the gains from selecting teachers based on their estimated VA. We begin by evaluating Hanushek's (2009) proposal to deselect the bottom 5% of teachers based on their value-added. We estimate that replacing a teacher whose true VA is in the bottom 5 percent with an average teacher would increase the present value of students' lifetime income by $267,000 per classroom taught.[8] However, because VA is estimated with noise, the gains from deselecting teachers based on a limited number of classrooms are smaller. We estimate the present value gains from deselecting the bottom 5% of teachers to be approximately $135,000 based on one year of data and $190,000 based on three years of data.

We then evaluate the expected gains from policies that pay bonuses to high-VA teachers in order to increase retention rates. The gains from such policies appear to be only modestly larger than their costs. Although the present value benefit from retaining a teacher whose estimated VA is at the 95th percentile after three years is nearly $200,000 per year, most bonus payments end up going to high-VA teachers who would have stayed even without the additional payment (Clotfelter et al. 2008). Replacing low VA teachers may therefore be a more cost effective strategy to increase teacher quality in the short run than paying bonuses to retain high-VA teachers. In the long run, higher salaries could attract more high VA teachers to the teaching profession, a potentially important benefit that we do not measure here.[9]

It is important to keep two caveats in mind when evaluating the policy implications of our findings. First, teachers were not incentivized based on test scores in the school district and time period we study. The signal content of value-added might be lower when it is used to evaluate teachers because of behavioral responses such as cheating or teaching to the test (Jacob and Levitt 2003, Jacob 2005, Neal and Schanzenbach 2010). Our results quantify the gains from higher VA teachers in an environment without such distortions in teacher behavior.[10] Further work is

[8] This calculation discounts the earnings gains at a rate of 5% to age 12. The total undiscounted earnings gains from this policy are $52,000 per child and more than $1.4 milllion for the average classroom.

[9] Increasing salaries or paying bonuses based on VA could also result in gains to students via changes in teacher effort in the short run. However, a recent experimental study from the U.S. found no significant impacts of this type of incentive program (Springer et al. 2010).

[10] Even in our sample, we find that the top 2% of teachers ranked by VA have patterns of test score gains that are consistent with test manipulation based on the proxy developed by Jacob and Levitt (2003). Correspondingly, these

needed to determine how VA should be used for education policy in a high stakes environment with multitasking and imperfect monitoring (Holmstrom and Milgrom 1991, Barlevy and Neal 2012).

Second, our analysis does not compare value-added with other measures of teacher quality. It is quite plausible that aspects of teacher quality which are not captured by standardized tests have significant long-term impacts. This raises the possibility that other measures of teacher quality (e.g., evaluations based on classroom observation) might be even better predictors of teachers' long-term impacts than value-added scores, though the signal content of these measures in a high stakes environment could also be degraded by behavioral distortions. Further work comparing the long-term impacts of teachers rated on various metrics is needed to determine the optimal method of teacher evaluation. What is clear from this study is that improving teacher quality is likely to yield substantial returns for students; the best way to accomplish that goal is less clear.

The paper is organized as follows. In Section 2, we present a statistical model to formalize the questions we seek to answer and derive estimating equations for our empirical analysis. Section 3 describes the data sources and provides summary statistics as well as cross-sectional correlations between scores and adult outcomes as a benchmark. Section 4 discusses the results of our tests for bias in VA measures. Results on teachers' long-term impacts are given in section 5. Section 6 presents policy calculations and Section 7 concludes.

2 Conceptual Framework

We structure our analysis using a stylized dynamic model of the education production function based on previous work (Todd and Wolpin 2003, Cunha and Heckman 2010, Cunha, Heckman, and Schennach 2010). The purpose of the model is to formalize the identification assumptions underlying our empirical analysis and clarify how the reduced-form parameters we estimate should be interpreted. We therefore focus exclusively on the role of teachers, abstracting from other inputs to the education production function, such as peers or parental investment. Using this model, we (1) define a set of reduced-form treatment effects, (2) present the assumptions under which we can identify these treatment effects, and (3) derive estimating equations for these parameters.

2.1 Structural Model of Student Outcomes

Our model is characterized by three relationships: a specification for test scores, a specification for earnings (or other adult outcomes), and a rule that governs student and teacher assignment to

high VA outlier teachers also have much smaller long-term impacts than one would predict based on their VA.

classrooms. School principals first assign student i in grade g to a classroom $c(i,g)$ based on lagged test scores, prior inputs, and other unobserved determinants of student achievement. Principals then assign a teacher j to each classroom c based on classroom characteristics such as mean lagged scores and class demographics. Let $j(i,g) = j(c(i,g))$ denote student i's teacher in grade g. Let e_j denote teacher j's years of teaching experience.

Student i's test score in grade g, A_{ig}, is a function of current and prior inputs:

$$A_{ig} = \sum_{s=1}^{g} \sigma_{sg}\mu_{j(i,s)} + \lambda_{c(i,g)} + \eta_i + \zeta_{ig} \tag{1}$$

where $\mu_{j(i,g)}$ represents the impact of teacher j on test scores, which we term the teacher's "value-added." We scale teacher quality so that the average teacher has quality $\mu_j = 0$ and the effect of teacher quality in grade g on scores in grade g is $\sigma_{gg} = 1$. For $s < g$, σ_{sg} measures the persistent impact of teacher quality μ in grade s on test scores at the end of grade g. $\lambda_{c(i,g)}$ represents an exogenous transitory classroom-level shock, η_i represents academic ability, and ζ_{ig} represents idiosyncratic noise and other period-specific innovations in individual achievement.

The model for scores in (1) makes two substantive restrictions that are standard in the value-added literature. First, it assumes that teacher quality μ_j is fixed over time, except for the effects of teacher experience, which we model in our empirical specifications. This rules out the possibility that teacher quality fluctuates across years (independent of experience) or that it depends upon the characteristics of the students assigned to the teacher (e.g., high vs low achieving students).[11] Second, our model does not explicitly account for endogenous responses of other inputs such as parental effort in response to changes in teacher quality. We discuss the consequences of these assumptions for our results below.

Earnings Y_i are a function of the inputs over all G grades:

$$Y_i = \sum_{g=1}^{G} \gamma_g \tau^Y_{j(i,g)} + \eta^Y_i \tag{2}$$

where $\tau^Y_{j(i,g)}$ represents teacher j's impact on earnings, γ_g measures the effect of teacher quality in grade g on earnings and η^Y_i reflects individual heterogeneity in earnings ability, which may be correlated with academic ability η_i. This specification assumes that the transitory classroom and individual-level shocks that affect scores have no impact on earnings, a simplification that has no effect on the results below.

[11] One could reinterpret λ in equation 1 as a class-specific component of teacher quality. In that case, the methods we implement below would estimate the component of teacher quality that is constant across years.

2.2 Identifying Teachers' Impacts on Scores

Our first goal is to identify the causal impacts of changing the teacher of class c from teacher j to j' in grade g on test scores and earnings. Define the potential outcome $A_{ig}(j')$ as the test score student i would have in grade g if his teacher were $j(i,g) = j'$. With the normalization $\sigma_{gg} = 1$, the causal effect of replacing teacher j with j' on student i's end-of-year score is simply $A_{ig}(j') - A_{ig}(j) = \mu_{j'} - \mu_j$. In our stylized model, the treatment effect $A_{ig}(j') - A_{ig}(j)$ coincides with the structural impact of teachers on scores. In a more general model with endogenous parent inputs and peer quality, this reduced-form treatment effect combines various structural parameters. For instance, students assigned to a better teacher may get less help on their homework from parents. Though it is not a policy-invariant primitive parameter, the reduced-form parameter μ_j is of direct relevance to certain questions, such as the impacts of retaining teachers on the basis of their VA (Todd and Wolpin 2003).

To estimate μ_j, we begin by estimating the following empirical model for student i's test score in grade g in school year t:

$$\begin{aligned} A_{igt} &= f_{1g}(A_{i,t-1}) + f_2(e_{j(i,g,t)}) + \phi_1 X_{igt} + \phi_2 \bar{X}_{c(i,g,t)} + \nu_{igt} \\ \text{where } \nu_{igt} &= \mu_{j(i,g,t)} + \theta_{c(i,g,t)} + \varepsilon_{igt} \end{aligned} \tag{3}$$

Here $f_{1g}(A_{i,t-1})$ is a control function for individual test scores in year $t-1$, $f_2(e_{j(i,g,t)})$ controls for the impacts of teacher experience, X_{igt} is a vector of student characteristics (such as whether the student is a native English speaker), and $\bar{X}_{c(i,g,t)}$ is a vector of classroom-level characteristics determined before teacher assignment (such as class size or an indicator for being an honors class). We decompose the error term in the empirical model into three components: teacher quality (μ_j), class shocks ($\theta_{c(i,g,t)}$), and idiosyncratic shocks (ε_{igt}). We can distinguish teacher effects μ_j from class shocks $\theta_{c(i,g,t)}$ by observing teachers over many school years.[12] Note that because we control for the effects of teacher experience in (3), μ_j represents the variation in teacher quality that is independent of experience.[13]

The empirical model for test scores in (3) differs from the structural model in (1) because we cannot observe all the terms in (1), such as heterogeneity in individual ability (η_i and ζ_{ig}). Value-

[12] This is the key distinction between our paper and Chetty et al.'s (2011) analysis of the long term impacts of Project STAR using tax data. Chetty et al. observe each teacher in only one classroom and therefore cannot separate teacher and class effects.

[13] To simplify notation, we assume that teachers teach one class per year (as in elementary schools). Because the j and c subscripts become redundant, we drop the c subscript. When teachers are assigned more than one class per year, we treat each class as if it were in a separate year for the purposes of the derivation below.

added models address this problem by controlling for prior-year test scores, which in principle should capture much of the variance in ability because η_i is a component of previous test scores. With these controls, the idiosyncratic error term in the empirical model ε_{igt} reflects unobserved student-level heterogeneity in test scores arising from the components of the structural model in (1) that are orthogonal to lagged scores and other observable characteristics. The class-level error term $\theta_{c(i,g,t)}$ reflects analogous unobserved class-level heterogeneity.

There are various methods one could use to estimate μ_j and the other error components in (3), such as estimating a correlated random effects model, a hierarchical linear model, or implementing an Empirical Bayes procedure. All of these methods rely on the following identification assumption to obtain consistent estimates of μ_j.

Assumption 1 Students are not sorted to teachers on unobservable determinants of test scores:

$$\mathbb{E}\left[\theta_{c(i,g,t)} + \varepsilon_{igt} | j\right] = \mathbb{E}\left[\theta_{c(i,g,t)} + \varepsilon_{igt}\right]$$

Assumption 1 requires that each teacher is no more likely than other teachers to be assigned students who score highly, conditional on the controls in the empirical model (3). If this assumption fails, the estimated teacher effects $\hat{\mu}_j$ will pick up differences in unobserved student characteristics across teachers and not the causal impacts of the teachers themselves. Note that Assumption 1 is not inconsistent with some parents sorting their children to particular teachers. Assumption 1 only requires that the observable characteristics $\left\{A_{i,t-1}, X_{igt}, \bar{X}_{c(i,g,t)}\right\}$ are sufficiently rich so that any remaining unobserved heterogeneity in test scores is balanced across teachers.[14] The first half of our empirical analysis focuses on assessing whether this is the case using two tests that we describe in Section 4.

Empirical Implementation. We estimate μ_j using an Empirical Bayes procedure following Morris (1983) and Kane and Staiger (2008, pp 14-16), which is the most commonly used approach to estimate VA (McCaffrey et al. 2003). We use this approach because of its computational simplicity and because our primary goal is to evaluate the properties of existing VA measures rather than devise new measures. Our procedure for estimating μ_j consists of three steps, which we implement separately for math and English observations:

Step 1: Calculate residual test score gains. We estimate (3) using OLS and compute residuals of student test scores, $\hat{\nu}_{igt}$. We then estimate the variances of the error components σ^2_μ, σ^2_θ, and

[14] For example, suppose motivated parents are able to get their children better teachers. These children would presumably also have had higher test scores in the previous grade. Hence, conditional on prior test scores, the remaining variation in current test scores could be balanced across teachers despite unconditional sorting.

σ_ε^2 using equations (2)-(4) in Kane and Staiger (2008). Intuitively, the within-classroom variance identifies σ_ε^2, the within-teacher cross-classroom covariance identifies σ_μ^2, and the remaining variance is due to σ_θ^2.

Step 2: Calculate average teacher effects. Let $\overline{\nu}_{jt}$ denote the mean score residual for the classroom taught by teacher j in year t and n_{jt} the number of students in that class. We estimate each teacher's quality using a precision-weighted average of $\overline{\nu}_{jt}$ across the classes taught by teacher j:

$$\bar{\nu}_j = \sum_t h_{jt}\bar{\nu}_{jt} / \sum_t h_{jt}$$

where $h_{jt} = 1/(\hat{\sigma}_\theta^2 + \hat{\sigma}_\varepsilon^2/n_{jt})$ denotes is the inverse of the variance of the estimate of teacher quality obtained from class t.

Step 3: Shrink teacher effect estimates. Finally, we shrink the mean test score impact $\overline{\nu}_j$ toward the sample mean (0) to obtain an estimate of the teacher's quality:

$$\hat{\mu}_j = \overline{\nu}_j \frac{\hat{\sigma}_\mu^2}{\hat{\sigma}_\mu^2 + 1/\sum_t h_{jt}} = \overline{\nu}_j \cdot r \tag{4}$$

where $r \equiv \frac{Var(\mu_j)}{Var(\overline{\nu}_j)}$ is commonly termed the "reliability" of the VA estimate.

To understand the purpose of the shrinkage correction, consider an experiment in which we estimate teacher impacts $\overline{\nu}_j$ in year t and then randomly assign students to teachers in year $t+1$. The best (mean-squared error minimizing) linear predictor of student's test scores $A_{ig,t+1}$ based on $\bar{\nu}_j$ is obtained from the OLS regression $A_{ig,t+1} = a + b\bar{\nu}_j$. The coefficients in this regression are $a = 0$ and $b = \frac{cov(A_{ig,t+1},\bar{\nu}_j)}{var(\overline{\nu}_j)} = \frac{Var(\mu_j)}{Var(\overline{\nu}_j)} = r$, implying that the optimal forecast of teacher j's impact on future scores is $\hat{\mu}_j = \overline{\nu}_j \cdot r$. From a frequentist perspective, the measurement error in $\overline{\nu}_j$ makes it optimal to used a biased but more precise estimate of teacher quality to minimize the mean-squared error of the forecast. From a Bayesian perspective, the posterior mean of the distribution of μ_j with Normal errors is a precision-weighted average of the sample mean ($\overline{\nu}_j$) and the mean of the prior (0), which is $\mathbb{E}\mu_j|\overline{\nu}_j = \overline{\nu}_j \cdot r = \hat{\mu}_j$. Because of these reasons, we follow the literature and use $\hat{\mu}_j$ as our primary measure of teacher quality in our empirical analysis. As a robustness check, we replicate our main results using mean test score residuals ($\bar{\nu}_j$) and show that, as expected, the estimated impacts are attenuated by roughly the mean of the shrinkage factor r.

2.3 Identifying Teachers' Impacts on Earnings

The impact of changing the teacher of class c from j to j' in grade g on mean earnings is:

$$\mu_j^Y - \mu_{j'}^Y = \mathbb{E}Y_i(j(i,g)) - \mathbb{E}Y_i(j'(i,g)) \tag{5}$$

$$= \gamma_g \left(\tau_{j'(i,g)}^Y - \tau_{j(i,g)}^Y\right) + \sum_{s=g+1}^{G} \gamma_s \left(\mathbb{E}\tau_{j(i,s)}^Y|_{j'(i,g)} - \mathbb{E}\tau_{j(i,s)}^Y|_{j(i,g)}\right). \tag{6}$$

Replacing teacher j affects earnings through two channels. The first term in (5) represents the direct impact of the change in teachers on earnings. The second term represents the indirect impact via changes in the expected quality of subsequent teachers to which the student is assigned. For example, a higher achieving student may be tracked into a more advanced sequence of classes taught by higher quality teachers. In a more general model, other determinants of earnings such as parental effort or peer quality might also respond endogenously to the change in teachers.

In principle, one could estimate teacher j's reduced-form causal impact on earnings, μ_j^Y, using an empirical model analogous to the one used above for test scores:

$$Y_i = f_{1g}^Y(A_{i,t-1}) + f_2^Y(e_{j(i,g,t)}) + \phi_1^Y X_{igt} + \phi_2^Y \bar{X}_{c(i,g,t)} + \nu_{igt}^Y \tag{7}$$

$$\nu_{igt}^Y = \mu_{j(i,g,t)}^Y + \theta_{c(i,g,t)}^Y + \varepsilon_{igt}^Y$$

Teacher impacts on earnings μ_j^Y can be identified under an assumption about sorting analogous to Assumption 1:

$$\mathbb{E}\left[\theta_{c(i,g,t)}^Y + \varepsilon_{igt}^Y \mid j\right] = \mathbb{E}\left[\theta_{c(i,g,t)}^Y + \varepsilon_{igt}^Y\right] \tag{8}$$

This condition, although similar to Assumption 1, is a much stronger requirement in practice. Assumption 1 holds if ε_{igt} is balanced across teachers, which requires that η_i is orthogonal to A_{igt} conditional on lagged test scores and other observables. The condition in (8) holds if ε_{ijt}^Y is balanced across teachers, which requires η_i^Y to be orthogonal to Y_i conditional on lagged test scores and other observables. Because η_i appears directly in $A_{i,t-1}$, it is likely to be absorbed by controlling for lagged scores. In contrast, η_i^Y does *not* appear in lagged scores and hence is unlikely to be absorbed by these controls. If we observed an analog of lagged scores such as lagged expected earnings, we could effectively control for η_i^Y and more plausibly satisfy (8).

As a concrete example, suppose that students have heterogeneous levels of ability, which affects scores and earnings, and family connections, which only affect earnings. Students are sorted to

teachers on the basis of both of these characteristics. While ability is picked up by lagged test scores and thus eliminated from ε_{igt}, family connections are not absorbed by the controls and appear in ε^Y_{igt}. As a result, teachers' impacts on scores can be consistently estimated, but their impacts on earnings cannot because there is systematic variation across teachers in their students' earnings due purely to connections.

In practice, we are unable to account for η^Y_i fully: tests for sorting on pre-determined characteristics analogous to those in Section 4.1 reveal that (8) is violated in our data. Therefore, we cannot identify teachers' total impacts on earnings μ^Y_j despite being able to identify teachers' impacts on test scores. Given this constraint, we pursue a less ambitious objective: estimating the correlation between teachers' impacts on scores and earnings, $cov(\mu_j, \mu^Y_j)$. This yields a lower bound on teacher effects on earnings μ^Y_j, as the standard deviation of μ^Y_j is bounded below by $\beta_g \sigma_\mu$, which measures the portion of $var(\mu^Y_j)$ due to $cov(\mu_j, \mu^Y_j)$.

To see how we can identify $cov(\mu_j, \mu^Y_j)$, consider the following empirical model for earnings as a function of teacher VA for student i in grade g in year t:

$$Y_i = \beta_g \hat{\mu}_{j(i,g)} + f^\mu_{1g}(A_{i,t-1}) + f^\mu_2(e_{j(i,g,t)}) + \phi^\mu_1 X_{igt} + \phi^\mu_2 \bar{X}_{c(i,g,t)} + \varepsilon^\mu_{igt}. \tag{9}$$

The coefficient β_g in this equation represents the mean increase in student earnings from a one unit increase in teacher VA in grade g, as measured using the Empirical Bayes procedure described above. Estimating (9) using OLS yields an unbiased estimate of β_g under the following assumption.

Assumption 2 Teacher value-added is orthogonal to unobserved determinants of earnings:

$$cov\left(\hat{\mu}_j, \varepsilon^\mu_{igt}\right) = 0.$$

Assumption 2 is weaker than (8) because it only requires that there be no correlation between teacher value-added and unobservables.[15] In our example above, it allows students with better family connections η^Y_i to be systematically tracked to certain teachers as long as those teachers do not systematically have higher levels of value-added on test scores, conditional on the controls $\left\{A_{i,t-1}, e_{j(i,g,t)}, X_{igt}, \bar{X}_{c(i,g,t)}\right\}$. While this remains a strong assumption, it may hold in practice because teacher VA was not publicized during the period we study and VA is very difficult to predict based on teacher observables. We evaluate whether conditioning on observables is adequate to

[15] Assumption 2 would be violated if the same observations were used to estimate $\widehat{\mu}_j$ and β because the estimation errors in (3) and (9) are correlated. Students with unobservably high test scores η_i are also likely to have unobservably high earnings η^Y_i. We deal with this technical problem by using a leave-out mean to estimate $\widehat{\mu}_j$ as described in Section 4.

satisfy Assumption 2 using quasi-experimental techniques in Section 5.

The coefficient β_g in (9) represents the reduced-form impact of having a higher VA teacher in grade g and includes the impacts of subsequent endogenous treatments such as better teachers in later grades. While this reduced-form impact is of interest to parents, one may also be interested in identifying the impact of each teacher net of potential tracking to better teachers in later grades. Let $\tilde{\beta}_g$ denote the impact of teacher VA in grade g on earnings holding fixed teacher VA in subsequent grades. One intuitive specification to identify $\tilde{\beta}_g$ is to regress earnings on teacher VA in all grades simultaneously:

$$Y_i = \sum_{g=1}^{G} \tilde{\beta}_g \hat{\mu}_{j(i,g)} + \varepsilon_i^{\mu}. \tag{10}$$

Identifying $\left\{\tilde{\beta}_g\right\}$ in (10) requires the orthogonality condition $Cov\left(\hat{\mu}_{j(i,g)}, \varepsilon_i^{\mu}\right) = 0$. As we discussed above, this assumption does not hold unconditionally because students are assigned to teachers in grade g based on grade $g-1$ test scores $A_{i,g-1}$. Because we must condition on $A_{i,g-1}$ in order to obtain variation in grade g teacher VA $\hat{\mu}_{j(i,g)}$ that is orthogonal to student characteristics, we cannot directly estimate (10), as $A_{i,g-1}$ is endogenous to grade $g-1$ teacher VA $\hat{\mu}_{j(i,g-1)}$.[16] Instead, we develop a simple iterative method of recovering the net impacts $\tilde{\beta}_g$ from our reduced form estimates β_g and estimates of the degree of teacher tracking in Section 6.1.

3 Data

We draw information from two databases: administrative school district records and information on these students and their parents from U.S. tax records. We first describe the two data sources and then the structure of the linked analysis dataset. Finally, we provide descriptive statistics and cross-sectional correlations using the analysis dataset.

3.1 School District Data

We obtain information on students, including enrollment history, test scores, and teacher assignments from the administrative records of a large urban school district. These data span the school years 1988-1989 through 2008-2009 and cover roughly 2.5 million children in grades 3-8. For simplicity, we refer below to school years by the year in which the spring term occurs (e.g., the school

[16] For the same reason, we also cannot estimate the complementarity of teachers across grades. Estimating complementarity requires simultaneous quasi-random assignment of teachers in *both* grades g and $g-1$, but we are only able to isolate quasi-random variation one grade at a time with our research design.

year 1988-89 is 1989).

Test Scores. The data include approximately 18 million test scores. Test scores are available for English language arts and math for students in grades 3-8 in every year from the spring of 1989 to 2009, with the exception of 7th grade English scores in 2002.[17] In the early and mid 1990s, all tests were specific to the district. Starting at the end of the 1990s, the tests in grades 4 and 8 were administered as part of a statewide testing system, and all tests in grades 3-8 became statewide in 2006 as required under the No Child Left Behind law.[18] Because of this variation in testing regimes, we follow prior work on measuring teachers' effects on student achievement, taking the official scale scores from each exam and normalizing the mean to zero and the standard deviation to one by year and grade. The within-grade variation in achievement in the district we examine is comparable to the within-grade variation nationwide, so our results can easily be compared to estimates from other samples.[19]

Demographics. The dataset contains information on ethnicity, gender, age, receipt of special education services, and limited English proficiency for the school years 1989 through 2009. The database used to code special education services and limited English proficiency changed in 1999, creating a break in these series that we account for in our analysis by interacting these two measures with a post-1999 indicator. Information on free and reduced price lunch is available starting in school year 1999.

Teachers. The dataset links students in grades 3-8 to classrooms and teachers from 1991 through 2009.[20] This information is derived from a data management system which was phased in over the early 1990s, so not all schools are included in the first few years of our sample. In addition, data on course teachers for middle and junior high school students—who, unlike students in elementary schools, are assigned different teachers for math and English—are more limited.

[17] We also have data on math and English test scores in grade 2 from 1991-1994 and English test scores in grades 9-10 from 1991-1993, which we use only when estimating teachers' impacts on past and future test scores. Because these observations are a very small fraction of our analysis sample, excluding them has little impact on the placebo tests and fade-out estimates reported in Figure 2.

[18] All tests were administered in late April or May during the early-mid 1990s, and students were typically tested in all grades on the same day throughout the district. Statewide testing dates varied to a greater extent, and were sometimes given earlier in the school year (e.g., February) during the latter years of our data.

[19] The standard deviation of 4th and 8th grade English and math achievement in this district ranges from roughly 95 percent to 105 percent of the national standard deviation on the National Assessment of Educational Progress, based on data from 2003 and 2009, the earliest and most recent years for which NAEP data are available. Mean scores are significantly lower than the national average, as expected given the urban setting of the district.

[20] 5% of students switch classrooms or schools in the middle of a school year. We assign these students to the classrooms in which they took the test to obtain an analysis dataset with one observation per student-year-subject. However, when defining class and school-level means of student characteristics (such as fraction eligible for free lunch), we account for such switching by weighting students by the fraction of the year they spent in that class or school.

Course teacher data are unavailable prior to the school year 1994, then grow in coverage to roughly 60% by school year 1998 and 85% by 2003. Even in the most recent years of the data, roughly 15 percent of the district's students in grades 6 to 8 are not linked to math and English teachers because some middle and junior high schools still do not report course teacher data.

The missing teacher links raise two potential concerns. First, our estimates (especially for grades 6-8) apply to a subset of schools with more complete information reporting systems and thus may not be representative of the district as a whole. Reassuringly, we find that these schools do not differ significantly from the sample as a whole on test scores and other observables. Second, and more importantly, missing data could generate biased estimates. Almost all variation in missing data occurs at the school level because data availability is determined by whether the school utilizes in the district's centralized data management system for tracking course enrollment and teacher assignment. Specifications that exploit purely within-school comparisons are therefore essentially unaffected by missing data and we show that our results are robust to exploiting such variation. Moreover, we obtain similar results for the subset of years when we have complete data coverage in grades 3-5, confirming that missing data does not drive our results.

We obtain information on teacher experience from human resource records. The human resource records track teachers since they started working in the district and hence give us an uncensored measure of within-district experience for the teachers in our sample. However, we lack information on teaching experience outside of the school district.

Sample Restrictions. Starting from the raw dataset, we make a series of sample restrictions that parallel those in prior work to obtain our primary school district sample. First, because our estimates of teacher value-added always condition on prior test scores, we restrict our sample to grades 4-8, where prior test scores are available. Second, we drop the 2% of observations where the student is listed as receiving instruction at home, in a hospital, or in a school serving solely disabled students. We also exclude the 6% of observations in classrooms where more than 25 percent of students are receiving special education services, as these classrooms may be taught by multiple teachers or have other special teaching arrangements. Third, we drop classrooms with less than 10 students or more than 50 students as well as teachers linked with more than 200 students in a single grade, because such students are likely to be mis-linked to classrooms or teachers (0.5% of observations). Finally, when a teacher is linked to students in multiple schools during the same year, which occurs for 0.3% of observations, we use only the links for the school where the teacher is listed as working according to human resources records and set the teacher as missing in the other

schools. After these restrictions, we are left with 15.0 million student-year-subject observations. Of these, 9.1 million records have information on teacher and 7.7 million have information on both teachers and test score gains, which we need to estimate value-added.

3.2 Tax Data

In Chetty, Friedman, and Rockoff (2011), we obtain data on students' adult outcomes and their parents' characteristics from income tax returns. Here, we briefly summarize some key features of the variables used in the analysis below. The year always refers to the tax year (i.e., the calendar year in which the income is earned or the college expense incurred). In most cases, tax returns for tax year t are filed during the calendar year $t+1$. We express all monetary variables in 2010 dollars, adjusting for inflation using the Consumer Price Index.

Earnings. Individual earnings data come from W-2 forms, which are available from 1999-2010. W-2 data are available for *both* tax filers and non-filers, eliminating concerns about missing data. Individuals with no W-2 are coded as having 0 earnings.[21] We cap earnings in each year at $100,000 to reduce the influence of outliers; 1.2% of individuals in the sample report earnings above $100,000 at age 28.

College Attendance. We define college attendance as an indicator for having one or more 1098-T forms filed on one's behalf. Title IV institutions – all colleges and universities as well as vocational schools and other postsecondary institutions – are required to file 1098-T forms that report tuition payments or scholarships received for every student. Because the 1098-T forms are filed directly by colleges, missing data concerns are minimal.[22] Comparisons to other data sources indicate that 1098-T forms accurately capture US college enrollment.[23] We have no information about college completion or degree attainment because the data are based on tuition payments. The 1098-T data are available from 1999-2009.

College Quality. We construct an earnings-based index of college quality as in Chetty et al. (2011). Using the full population of all individuals in the United States aged 20 on 12/31/1999

[21] We obtain similar results using household adjusted gross income reported on individual tax returns. We focus on the W-2 measure because it provides a consistent definition of individual wage earnings for both filers and non-filers. One limitation of the W-2 measure is that it does not include self-employment income.

[22] Colleges are not required to file 1098-T forms for students whose qualified tuition and related expenses are waived or paid entirely with scholarships or grants; however, the forms are generally available even for such cases, perhaps because of automated reporting to the IRS by universities.

[23] See Chetty et al. (2011) for a comparison of total enrollment based on 1098-T forms and statistics from the Current Population Survey. Chetty et al. use this measure to analyze the impacts of Project STAR on college attendance. Dynarski et al. (2011) show that using data on college attendance from the National Clearinghouse yields very similar estimates to Chetty et al.'s findings, providing further confirmation that the 1098-T based college indicator is accurate.

and all 1098-T forms for year 1999, we group individuals by the higher education institution they attended in 1999. We take a 0.25% random sample of those not attending a higher education institution in 1999 and pool them together in a separate "no college" category. For each college or university (including the "no college" group), we then compute average W-2 earnings of the students in 2009 when they are aged 30. Among colleges attended by students in our data, the average value of our earnings index is $42,932 for four-year colleges and $28,093 for two-year colleges.[24] For students who did not attend college, the imputed mean earnings level is $16,361.

Neighborhood Quality. We use data from 1040 forms to identify each household's ZIP code of residence in each year. For non-filers, we use the ZIP code of the address to which the W-2 form was mailed. If an individual did not file and has no W-2 in a given year, we impute current ZIP code as the last observed ZIP code. We construct a measure of a neighborhood's SES using data on the percentage of college graduates in the individual's ZIP code from the 2000 Census.

Retirement Savings. We measure retirement savings using contributions to 401(k) accounts reported on W-2 forms from 1999-2010. We define saving for retirement as an indicator for ever contributing to a 401(k) during this period.

Teenage Birth. We first identify all women who claim a dependent when filing their taxes at any point before the end of the sample in tax year 2010. We observe dates of birth and death for all dependents and tax filers until the end of 2010 as recorded by the Social Security Administration. We use this information to identify women who ever claim a dependent who was born while the mother was a teenager (between the ages of 13 and 19 as of 12/31 the year the child was born). We refer to this outcome as having a "teenage birth," but note that this outcome differs from a direct measure of teenage birth in three ways. First, it does not capture teenage births to individuals who never file a tax return before 2010. Second, the mother must herself claim the child as a dependent at some point during the sample years. If the child is claimed as a dependent by the grandmother for all years of our sample, we would never identify the child. In addition to these two forms of under-counting, we also over-count the number of children because our definition could miscategorize other dependents as biological children. Because most such dependents tend to be elderly parents, the fraction of cases that are incorrectly categorized as teenage births is likely to be small. Even though this variable does not directly measure teenage births, we believe that it is a useful measure of outcomes in adulthood because it correlates with observables as expected (see

[24] For the small fraction of students who attend more than one college in a single year, we define college quality based on the college that received the largest tuition payments on behalf of the student.

Section 5.3). For instance, women who score higher on tests, attend college, or have higher income parents are significantly less likely to have teenage births.

Parent Characteristics. We link students to their parents by finding the earliest 1040 form from 1996-2010 on which the student was claimed as a dependent. We identify parents for 94.7% of students linked with tax records as adults. The remaining students are likely to have parents who did not file tax returns in the early years of the sample when they could have claimed their child as a dependent, making it impossible to link the children to their parents. Note that this definition of parents is based on who claims the child as a dependent, and thus may not reflect the biological parent of the child.

We define parental household income as Adjusted Gross Income (capped at $117,000, the 95th percentile in our sample), averaged over the three years when the child was 19-21 years old.[25] For years in which parents did not file, we impute parental household income from wages and unemployment benefits, each of which are reported on third-party information forms. We define marital status, home ownership, and 401(k) saving as indicators for whether the first primary filer who claims the child ever files a joint tax return, makes a mortgage interest payment (based on data from 1040's for filers and 1099's for non-filers), or makes a 401(k) contribution (based on data from W-2's) during the years when the child is between 19 and 21. We define mother's age at child's birth using data from Social Security Administration records on birth dates for parents and children. For single parents, we define the mother's age at child's birth using the age of the filer who claimed the child, who is typically the mother but is sometimes the father or another relative.[26] When a child cannot be matched to a parent, we define all parental characteristics as zero, and we always include a dummy for missing parents in regressions that include parent characteristics.

3.3 Analysis Dataset

Because most of the adult outcomes we analyze are at age 20 or afterward, we restrict our linked analysis sample to students who would graduate high school in the 2007-08 school year (and thus

[25] To account for changes in marital status, we always follow the primary filer who first claimed the child and define parent characteristics based on the tax returns filed by that parent when the child is between 19 and 21. For instance, if a single mother has a child and gets married when the child was 18, we would define household income as AGI including the mother and her new husband when the child is 19-21. If the child does not turn 21 before 2010, we code the parent characteristics as missing.

[26] We define the mother's age at child's birth as missing for 471 observations in which the implied mother's age at birth based on the claiming parent's date of birth is below 13 or above 65. These are typically cases where the parent does not have an accurate birth date recorded in the SSA file.

turn 20 in 2010) if they progress through school at a normal pace.[27] The school district records were linked to the tax data using an algorithm based on standard identifiers (date of birth, state of birth, gender, and names) described in Appendix A, after which individual identifiers were removed to protect confidentiality. 89.2% of the observations in the school district data were matched to the tax data and match rates do not vary with teacher VA (see Table 2 below).

The linked analysis dataset has one row per student per subject (math or English) per school year, as illustrated in Appendix Table 1. Each observation in the analysis dataset contains the student's test score in the relevant subject test, demographic information, and class and teacher assignment if available. Each row also lists all the students' available adult outcomes (e.g. college attendance and earnings at each age) as well as parent characteristics. We organize the data in this format so that each row contains information on a treatment by a single teacher conditional on pre-determined characteristics, facilitating estimation of equation (3). We account for the fact that each student appears multiple times in the dataset by clustering standard errors as described in section 4.1.

To maximize precision, we estimate teacher value-added using all years for which school district data are available (1991-2009). However, the impacts of teacher VA on test scores and adult outcomes that we report in the main text use only the observations in the linked analysis dataset (i.e., exclude students who would graduate high school after 2008), unless otherwise noted.[28]

3.4 Summary Statistics

The analysis dataset contains 6.0 million student-year-subject observations, of which 4.8 million have information on teachers. Table 1 reports summary statistics for the linked analysis dataset; see Appendix Table 2 for corresponding summary statistics for the full school district data used to estimate teacher value-added. Note that the summary statistics are student-school year-subject means and thus weight students who are in the district for a longer period of time more heavily, as does our empirical analysis. There are 974,686 unique students in our analysis dataset; on average, each student has 6.14 subject-school year observations.

The mean test score in the analysis sample is positive and has a standard deviation below

[27] A few classrooms contain students at different grade levels because of retentions or split-level classroom structures. To avoid dropping a subset of students within a classroom, we include every classroom that has at least one student who would graduate school during or before 2007-08 if he progressed at the normal pace. That is, we include all classrooms in which $\min_i(12+ \text{ school year} - \text{grade}_i) \leq 2008$.

[28] Within the analysis data, we use all observations for which the necessary data are available. In particular, when estimating the impacts of VA on scores, we include observations that were not matched to the tax data.

1 because we normalize the test scores in the full population that includes students in special education classrooms and schools (who typically have lower test scores). The mean age at which students are observed is 11.7 years. 76% of students are eligible for free or reduced price lunches. 2.7% of the observations are for students who are repeating the current grade.

The availability of data on adult outcomes naturally varies across cohorts. There are more than 4.6 million observations for which we observe college attendance at age 20. We observe earnings at age 25 for 2.2 million observations and at age 28 for 850,000 observations. Because many of these observations at later ages are for older cohorts of students who were in middle school in the early 1990s, they do not contain information on teachers. As a result, there are only 1.4 million student-subject-school year observations for which we see *both* teacher assignment and earnings at age 25, 376,000 at age 28, and only 63,000 at age 30. The oldest age at which the sample is large enough to obtain reasonably precise estimates of teachers' impacts on earnings turns out to be age 28. Mean earnings at age 28 is $20,327 (in 2010 dollars), which includes zero earnings for 34% of the sample.

For students whom we are able to link to parents, mothers are 28 years old on average when the student was born. The mean parent household income is $35,476, while the median is $27,144 Though our sample includes more low income households than would a nationally representative sample, it still includes a substantial number of higher income households, allowing us to analyze the impacts of teachers across a broad range of the income distribution. The standard deviation of parent income is $31,080, with 10% of parents earning more than $82,630.

As a benchmark for evaluating the magnitude of the causal effects estimated below, Appendix Tables 3-6 report estimates of OLS regressions of the adult outcomes we study on test scores. Both math and English test scores are highly positively correlated with earnings, college attendance, and neighborhood quality and are negatively correlated with teenage births. In the cross-section, a 1 SD increase in test score is associated with a $7,440 (37%) increase in earnings at age 28. Conditional on prior-year test scores and other controls that we use in our analysis below, a 1 SD increase in the current test score is associated with $2,545 (11.6%) increase in earnings on average. We show below that the causal impact of teacher VA on earnings is commensurate to this correlation in magnitude.

4 Does Value-Added Accurately Measure Teacher Quality?

Recent studies by Kane and Staiger (2008) and Rothstein (2010) among others have reached conflicting conclusions about whether VA estimates are biased by student sorting (i.e., whether Assumption 1 in Section 2.2 holds). In this section, we revisit this debate by presenting new tests for bias in VA estimates.

4.1 Empirical Methodology

Throughout our empirical analysis, we regress various outcomes on estimated teacher value-added. In this subsection, we discuss four aspects of our methodology that are relevant for all the regression estimates reported below: (1) leave-out mean estimation of VA, (2) control vectors, (3) standard error calculations, and (4) the treatment of outliers.

First, there is a mechanical correlation between $\hat{\mu}_j$ and student outcomes in a given school year because $\hat{\mu}_j$ is estimated with error and these errors also affect student outcomes.[29] We address this problem by following Jacob, Lefgren and Sims (2010) and use a leave-year-out (jackknife) mean to calculate teacher quality.[30] For example, when predicting teachers' effects on student outcomes in 1995, we estimate $\widehat{\mu}_j^{1995}$ based on all years of the sample *except* 1995. We then regress outcomes for students in 1995 on $\widehat{\mu}_j^{1995}$. More generally, for each observation in year t, we omit score residuals from year t when calculating teacher quality.[31] This procedure is essential to eliminate mechanical biases due to estimation error in $\hat{\mu}_j$ both in our tests for sorting and our estimates of teachers' impacts on adult outcomes.[32]

Second, we use a control vector that parallels existing VA models (e.g., Kane and Staiger 2008)

[29] This problem does not arise when estimating the impacts of treatments such as class size because the treatment is observed; here, the size of the treatment (teacher VA) must itself be estimated, leading to correlated estimation errors.

[30] Because we need at least two classes to define a leave-out mean, our analysis only applies to the population of teachers whom we see teaching two or more classes between 1991 and 2009. Among the classrooms with the requisite controls to estimate value-added (e.g. lagged test scores), we are unable to calculate a leave-out measure of VA for 9% of students because their teachers are observed in the data for only one year. The first-year VA of teachers who leave after one year is 0.01 SD lower than the first-year VA of those who stay for more years. Hence, the mean VA of the subset of teachers in our sample is only 0.001 SD higher than mean VA in the population, suggesting that our estimates are likely to be fairly representative of teacher effects in the full population.

[31] An alternative approach is to split the sample in two, for instance using data after 1995 to estimate teacher VA and data before 1995 to estimate its impacts on outcomes for students who are old enough to be seen in the tax data. We find that such a split-sample approach yields similar but less precise estimates.

[32] Regressing student outcomes on teacher VA without using a leave-out mean effectively introduces the same estimation errors on both the left and right hand side of the regression, yielding biased estimates of teachers' causal impacts. This is the reason that Rothstein (2010) finds that "fifth grade teachers whose students have above average fourth grade gains have systematically lower estimated value-added than teachers whose students underperformed in the prior year."

to estimate student test score residuals using (3):

$$A_{igt} = f_{1g}(A_{i,t-1}) + f_2(e_{j(i,g,t)}) + \phi_1 X_{igt} + \phi_2 \bar{X}_{c(i,g,t)} + \nu_{igt}$$

We parameterize the control function for lagged test scores $f_{1g}(A_{i,t-1})$ using a cubic polynomial in prior-year scores in math and a cubic in prior-year scores in English. We interact these cubics with the student's grade level to permit flexibility in the persistence of test scores as students age. We parametrize the control function for teacher experience $f_2(e_{j(i,g,t)})$ using dummies for years of experience from 0 to 5, with the omitted group being teachers with 6 or more years of experience.[33] The student-level control vector X_{igt} consists of the following variables: ethnicity, gender, age, lagged suspensions and absences, and indicators for grade repetition, special education, limited English. The class-level control vector $\bar{X}_{c(i,g,t)}$ includes (1) class size and class-type indicators (honors, remedial), (2) cubics in class and school-grade means of prior-year test scores in math and English each interacted with grade, (3) class and school-year means of all the individual covariates X_{igt}, and (4) grade and year dummies. To avoid estimating VA based on very few observations, we follow Kane and Staiger (2008) and exclude classrooms that have fewer than 7 observations with test scores and the full vector of controls X_{igt} (2% of observations). Importantly, the control vectors X_{igt} and $\bar{X}_{c(i,g,t)}$ consist entirely of variables from the school district dataset. We adopt this approach because our goal is to assess properties of value-added estimated without access to information available in tax data, which will not typically be available to school districts.

When estimating the impacts of teacher VA on adult outcomes using (9), we omit the student-level controls X_{igt}. By omitting X_{igt}, we can conduct most of our analysis of long-term impacts using a dataset collapsed to class means, which significantly reduces computational costs. We show in Appendix Table 7d that the inclusion of individual controls has little impact on the coefficients and standard errors of interest for a selected set of specifications.

Third, our outcomes have a correlated error structure because students within a classroom face common class-level shocks and because our analysis dataset contains repeat observations on students in different grades. One natural way to account for these two sources of correlated errors is to cluster standard errors by both student and classroom (Cameron, Gelbach, and Miller 2011). Unfortunately, implementing two-way clustering on a dataset with 6 million observations was infeasible because of computational constraints. We instead cluster standard errors at the

[33] We choose this functional form because prior work (e.g. Rockoff 2004) has shown that the impacts of teacher experience rise sharply and then stabilize after the first three years.

school by cohort level, which adjusts for correlated errors across classrooms and repeat student observations within a school. Clustering at the school-cohort level is convenient because it again allows us to conduct our analysis on a dataset collapsed to class means. We evaluate the robustness of our results to alternative forms of clustering in Appendix Table 7 and show that school-cohort clustering yields more conservative confidence intervals than the more computationally intensive techniques.

Finally, in our baseline specifications, we exclude classrooms taught by teachers whose estimated VA $\hat{\mu}_j^t$ falls in the top two percent for their subject (above 0.21 in math and 0.13 in English) because these teachers' impacts on test scores appear suspiciously consistent with testing irregularities indicative of cheating. Jacob and Levitt (2003) develop a proxy for cheating that measures the extent to which a teacher generates very large test score gains that are followed by very large test score losses for the same students in the subsequent grade. Jacob and Levitt establish that this is a valid proxy by showing that it is highly correlated with unusual answer sequences that directly point to test manipulation. Teachers in the top 2% of our estimated VA distribution are significantly more likely to show suspicious patterns of test scores gains, as defined by Jacob and Levitt's proxy (see Appendix Figure 1).[34] We therefore trim the top 2% of outliers in all the specifications reported in the main text. We investigate how trimming at other cutoffs affects our main results in Appendix Table 8. The qualitative conclusion that teacher VA has long-term impacts is not sensitive to trimming, but including teachers in the top 2% reduces our estimates of teachers' impacts on long-term outcomes by 20-40%. In contrast, excluding the bottom 2% of the VA distribution has little impact on our estimates, consistent with the view that test manipulation to obtain high test score gains is responsible for the results in the upper tail. Directly excluding teachers who have suspect classrooms based on Jacob and Levitt's proxy for cheating yields very similar results to trimming on VA itself.

Because we trim outliers, our baseline estimates should be interpreted as characterizing the relationship between VA and outcomes below the 98th percentiles of VA. This is the relevant range for many questions, such as calculating the gains of switching a child from an average teacher to a teacher 1 SD above the mean. If school districts can identify and eliminate teacher cheating –

[34] Appendix Figure 1 plots the fraction of classrooms that are in the top 5 percent according to Jacob and Levitt's proxy, defined in the notes to the figure, vs. our leave-out-year measure of teacher value-added. On average, classrooms in the top 5 percent according to the Jacob and Levitt measure have test score gains of 0.46 SD in year t followed by mean test score *losses* of 0.43 SD in the subsequent year. Stated differently, teachers' impacts on future test scores fade out much more rapidly in the very upper tail of the VA distribution. Consistent with this pattern, these exceptionally high VA teachers also have very little impact on their students' long-term outcomes.

e.g. by analyzing the persistence of test score gains as suggested by Jacob and Levitt – our estimates would also shed light on the gains from retaining the remaining high-VA teachers. Nevertheless, the fact that high-VA outliers do not have lasting impacts on scores or adult outcomes serves as a warning about the risks of manipulability of VA measures. The signal content of VA measures could be severely reduced if teachers game the system further when VA is actually used to evaluate teachers. This is perhaps the most important caveat to our results and a critical area for further work, as we discuss in the conclusion.

4.2 VA Estimates and Out-of-Sample Forecasts

The first step in our empirical analysis is to estimate leave-year-out teacher effects $\hat{\mu}_j^t$ for each teacher j and year t in our sample. We estimate VA using all years in the school district data for which we have teacher information (1991-2009). The standard deviation of teacher effects is $\sigma_\mu = 0.118$ in math and $\sigma_\mu = 0.081$ in English, very similar to estimates from prior work. Note that these standard deviations measure the dispersion in teacher effects that is orthogonal to teacher experience as well as other controls.[35] Throughout, we scale $\hat{\mu}_j^t$ in units of *student* test scores, i.e., a 1 unit increase in $\hat{\mu}_j^t$ refers to a teacher whose VA is predicted to raise student test scores by 1 SD. Because the standard deviation of teacher effects is approximately 0.1 SD of the student test score distribution (averaging across math and English), a 1 SD increase in teacher VA corresponds to an increase of 0.1 in $\hat{\mu}_j^t$.

We begin our evaluation of the properties of $\widehat{\mu}_j^t$ by verifying that our VA estimates have predictive power for test score gains outside the sample on which they were estimated. Under our assumption in (3) that true teacher effects μ_j are time-invariant, a 1 SD increase in $\widehat{\mu}_j^t$ should be associated with a 1 SD increase in test scores in year t.[36] Figure 1a plots student test scores (combining English and math observations) vs. our leave-year-out estimate of teacher VA in our linked analysis dataset. We condition on the classroom-level controls used when estimating the value-added model in this and all subsequent figures by regressing both the x- and y-axis variables on the vector of controls and then computing residuals. We then bin the student-subject-year residuals into twenty equal-size groups (vingtiles) of $\widehat{\mu}_j^t$ and plot the mean residual score in each

[35] Students assigned to first-year teachers have 0.03 SD lower test score gains, consistent with prior work. Because the impact of experience on scores is small, we have insufficient power to estimate its impacts on adult outcomes; we can rule out neither 0 effects nor effects commensurate to the impacts of VA estimated below. We therefore do not analyze teacher experience further in this paper.

[36] Although the estimation error in value-added leads to attenuation bias, the shrinkage correction we implement in (4) exactly offsets the attenuation bias so that a 1 unit increase in $\widehat{\mu}_j^t$ should raise scores by 1 unit.

bin. Note that these binned scatter plots provide a non-parametric representation of the conditional expectation function but do not show the underlying variance in the individual-level data. The regression coefficient and standard error reported in each figure are estimated on the micro data, with standard errors clustered by school-cohort as described above.

Figure 1a shows that a teacher with $\widehat{\mu}_j^t = 1$ generates a 0.86 SD increase in students' test scores in year t, with a t-statistic over 80 (see also Column 1 of Table 2). This confirms that the VA estimates are highly predictive of student test scores. The coefficient on $\widehat{\mu}_j^t$ is below 1, consistent with the findings of Kane and Staiger (2008), most likely because teacher value-added is not in fact a time-invariant characteristic. For instance, teacher quality may fluctuate when teachers switch schools or grades (Jackson 2010) and may drift over time for other reasons (Goldhaber and Hansen 2010). Such factors reduce the accuracy of forecasts based on data from other years. Because we estimate teacher VA using data from 1991-2009 but only include cohorts who graduate from high school before 2008 in our analysis dataset, the time span between the point at which we estimate VA and analyze test score impacts is especially large in our analysis sample. Replicating Column 1 of Table 2 on the full sample used to estimate teacher VA yields a coefficient on $\widehat{\mu}_j^t$ of 0.96. Because we are forced to use data from more distant years to identify value-added, our estimates of the impacts of teacher quality on adult outcomes may be slightly downward-biased.[37]

The relationship between $\widehat{\mu}_j^t$ and students' test scores in Figure 1a could reflect either the causal impact of teachers on achievement or persistent differences in student characteristics across teachers. For instance, $\widehat{\mu}_j^t$ may forecast students' test score gains in other years simply because some teachers are always assigned students with higher income parents. We now implement two sets of tests for such sorting.

4.3 Test 1: Selection on Observable Characteristics

Value-added estimates consistently measure teacher quality only if they are uncorrelated with unobserved components of student scores. A natural first test of this identifying assumption is to examine the correlation between our estimates of VA and variables omitted from standard VA models.[38] We use two sets of variables to evaluate selection: parent characteristics and prior test

[37] We do not account for variation over time in VA because our primary goal is to assess the properties of teacher VA measures currently being used by school districts. In future work, it would be interesting to develop time-varying measures of VA and evaluate whether they are better predictors of adult outcomes.

[38] Such correlation could arise from either actual selection of students to teachers with higher quality μ_j or sorting across teachers that is unrelated to true quality but generates measurement error in $\hat{\mu}_j^t$ that is correlated with student characteristics. Either of these sources of correlation would violate Assumption 1 and generate biased estimates of VA.

scores.

Parent Characteristics. The parent characteristics from the tax data are ideal to test for selection because they have not been used to fit value-added models in prior work but are strong predictors of student achievement. We collapse the parent characteristics into a single index by regressing test scores on mother's age at child's birth, indicators for parent's 401(k) contributions and home ownership, and an indicator for the parent's marital status interacted with a quartic in parent's household income.[39] Let A^p_{it} denote the predicted test score for student i in year t in this regression, which we calculate only for students for whom test score data are available. These predicted test scores are an average of the parent characteristics, weighted optimally to reflect their relative importance in predicting test scores. The standard deviation of predicted test scores is 0.26, roughly 30% of the standard deviation of actual test scores in our analysis sample.

Figure 1b plots $\widehat{A}^p_{c,g-1}$ against teacher VA measured using a leave-year-out mean as described above. There is no relationship between predicted scores and teacher VA. At the upper bound of the 95% confidence interval, a 1 standard deviation increase in teacher VA raises predicted scores based on parent characteristics by 0.01 SD (see also Column 2 of Table 2). This compares with an actual score impact of 0.86 SD, showing that very little of the association between teacher VA and actual test scores is driven by sorting on omitted parent characteristics. Note that this result does *not* imply that students from higher vs. lower socioeconomic status families uniformly get teachers of the same quality. Our finding is that controlling for the rich set of observables available in school district databases, such as test scores in the previous grade, is adequate to account for sorting of students to teachers based on parent characteristics. That is, if we take two students who have the same 4th grade test scores, classroom characteristics, ethnicity, suspensions, etc., the student assigned to a teacher with higher estimated VA in grade 5 does not systematically have different parental income or other characteristics.

A second, closely related method of assessing selection on parent characteristics is to control for predicted scores A^p_{it} when estimating the impact of VA on actual scores. Columns 3-4 in Table 2 restrict to the sample in which both score and predicted score are non-missing; the coefficient on $\widehat{\mu}^t_j$ changes only from 0.866 to 0.864 after controlling for predicted scores. Note that parent characteristics have considerable predictive power for test scores even conditional on the controls

[39] We code the parent characteristics as 0 for the 5.4% of matched students for whom we are unable to find a parent, and include an indicator for having no parent matched to the student. We also code mother's age at child's birth as 0 for a small number of observations where we match parents but do not have data on parents' ages, and include an indicator for such cases.

used to estimate the value-added model; the t-statistic on the predicted score A_{it}^{p} exceeds 60. The fact that parent characteristics are strong predictors of residual test scores yet are uncorrelated with $\widehat{\mu}_j^t$ suggests that the degree of bias in VA estimates is likely to be small (Altonji, Elder, and Taber 2005).

A third approach to evaluating the extent to which the omission of parent characteristics affects VA estimates is to re-estimate $\widehat{\mu}_j^t$, controlling for the parent characteristics to begin with. We repeat the three-step estimation procedure in Section 2.2, controlling for mean parent characteristics by classroom when estimating (3) using the same functional form used above to predict test scores. We then correlate estimates of teacher VA that control for parent characteristics with our original estimates that condition only on school-district observables. The correlation coefficient between the two VA estimates is 0.999, as shown in rows 1 and 2 of Table 3. All three tests show that selection on previously unobserved parent characteristics generates minimal bias in standard VA estimates.

Prior Test Scores. Another set of pre-determined variables that can be used to test for selection are prior test scores (Rothstein 2010). Because value-added models control for $A_{i,t-1}$, one can only evaluate sorting on $A_{i,t-2}$ (or, equivalently, on lagged gains, $A_{i,t-1} - A_{i,t-2}$). The question is whether controlling for additional lags substantially affects VA estimates once one controls for $A_{i,t-1}$. We now present three tests to answer this question that parallel those above for parent characteristics.

We first examine whether twice-lagged test scores are correlated with our baseline estimates of VA. Figure 1c plots twice-lagged scores $A_{i,t-2}$ against teacher VA, following the same methodology used to construct Figure 1a. There is virtually no relationship between VA and twice-lagged score conditional on the controls used to estimate the VA model. As a result, controlling for $A_{i,t-2}$ when estimating the impact of VA on out-of-sample test scores has little effect on the estimated coefficient (columns 6-7 of Table 2). The coefficient on VA is stable despite the fact that $A_{i,t-2}$ has significant predictive power for $A_{i,t}$, even conditional on $A_{i,t-1}$ and $\bar{X}_c$; the t-statistic on $A_{i,t-2}$ exceeds 350. Finally, controlling flexibly for $A_{i,t-2}$ at the individual level (using cubics in math and English scores) when estimating the VA model does not affect estimates significantly. The correlation coefficient between our baseline VA estimates and estimates that control for $A_{i,t-2}$ is 0.975, as shown in row 3 of Table 3. We conclude based on these tests that selection on grade $t-2$ scores generates minimal bias in VA estimates once one conditions on $t-1$ characteristics.

We further develop this test by examining the correlation of our baseline VA measure with

additional leads and lags of test scores. If our VA measures reflect the causal impact of teachers, the correlation between current teacher VA on test scores should jump in the current year. To test this hypothesis, we estimate (9), changing the dependent variable to test scores $A_{i,t+s}$ for $s \in [-4, 4]$, four years before and after the current grade t. Figure 2 plots the coefficients on current teacher VA from each of these regressions.[40] As predicted, teachers' impacts on scores jump at the end of the grade taught by that teacher. A 1 unit increase in teacher VA raises end-of-grade test scores by 0.86 SD, matching the estimate in column 1 of Table 2. In contrast, the same increase in teacher VA in grade g has essentially no impact on test scores prior to grade g. This finding suggests that VA measures capture causal effects of teachers rather than systematic differences across teachers in their students' characteristics, as such characteristics would have to be uncorrelated with past test scores and only affect the current score.

Figure 2 also shows that the impact of current teacher VA fades out in subsequent grades. Prior studies (e.g., Kane and Staiger 2008, Jacob, Sims, and Lefgren 2010, Rothstein 2010) document similar fade-out after one or two years but have not determined whether test score impacts continue to deteriorate after that point. The broader span of our dataset allows us to estimate test score persistence more precisely.[41] In our data, the impact of a 1 SD increase in teacher quality stabilizes at approximately 0.3 SD after 3 years, showing that students assigned to teachers with higher VA achieve long-lasting test score gains.

The last column of Table 2 analyzes the correlation between teacher VA and the probability that a student is matched to the tax data. In this column, we regress an indicator for being matched on teacher VA, using the same specification as in the other columns. There is no significant relationship between VA and match rates, suggesting that our estimates of the impacts of VA on outcomes in adulthood are unlikely to be biased by attrition.

4.4 Test 2: Teacher Switching Quasi-Experiments

The preceding tests show that the bias in VA estimates due to the omission of observables such as parent characteristics and twice-lagged scores is minimal. They do not, however, rule out the

[40] The estimates underlying this figure and their associated standard errors are reported in Appendix Table 9. Naturally, the grades used to estimate each of the points in Figure 2 vary because scores are only available for grades 3-8. We continue to find that VA has an effect on prior test scores that is two orders of magnitude smaller than its impact on current test scores if we restrict to individual grades and use the available leads and lags (e.g. two leads and two lags for grade 6).

[41] For instance, Jacob, Lefgren, and Sims estimate one-year persistence using 32,422 students and two-year persistence using 17,320 students. We estimate one-year persistence using more than 2.8 million student-year-subject observations and four-year persistence using more than 790,000 student-year-subject observations.

possibility that students are sorted to teachers based on unobservable characteristics orthogonal to these variables. The ideal method of testing for selection on unobservables is to evaluate whether VA estimates using observational data accurately predict students' test score gains when students are randomly assigned to teachers. Kane and Staiger (2008) implement such an experiment in Los Angeles involving approximately 3,500 students and 150 teachers. Kane and Staiger's point estimates suggest that there is little bias in VA estimates, but their 95% confidence interval is consistent with bias of up to 50% because of their relatively small sample size (Rothstein 2010). Moreover, Rothstein notes that because certain classes and schools were excluded from the experiment, the external validity of the findings is unclear.

Motivated by these concerns, we develop a quasi-experimental method of estimating the degree of bias due to selection on unobservables. Our approach yields more precise estimates of the degree of bias on a representative sample of a school district's student population.

Research Design. Our research design exploits the fact that adjacent cohorts of students within a school are frequently exposed to teachers with very different levels of VA because of teacher turnover. In our school district dataset, 14.5% of teachers switch to a different grade within the same school the following year, 6.2% of teachers switch to a different school within the same district, and another 6.2% switch out of the district entirely. These changes in the teaching staff from one year to the next generate variation in VA that is "quasi-experimental" in the sense that it is plausibly orthogonal to students' characteristics.

To understand our test, suppose a high-VA teacher moves from 4th grade in school s to another school between 1994 and 1995. Because students entering grade 4 in school s in 1995 have lower VA teachers on average, their mean test scores should be lower than the 1994 cohort if VA estimates capture teachers' causal impacts. Moreover, the size of the change in test scores across these adjacent cohorts should correspond to the change in mean VA. For example, in a school-grade cell with three classrooms, the loss of a math teacher with a VA estimate of 0.3 based on prior data should decrease average math test scores in the entire school-grade cell by 0.1. Importantly, because we analyze the data at the school-grade level, we do *not* exploit information on classroom assignment for this test, eliminating any bias due to non-random assignment of students across classrooms.

Changes in the quality of the teaching staff across school years constitute quasi-experimental variation under the assumption that they are uncorrelated with changes in the quality of students across adjacent cohorts. Let $\Delta\overline{\hat{\mu}}_{sgmt}$ denote the change in mean teacher VA $\hat{\mu}_{sgmt}$ from year

$t-1$ to year t in grade g in subject m (math or reading) in school s, and define mean changes in student unobservables $\Delta\overline{\varepsilon}_{sgmt}$ and $\Delta\overline{\varepsilon}^{\mu}_{sgmt}$ analogously. The identification assumption underlying the quasi-experimental design is

$$Cov\left(\Delta\overline{\hat{\mu}}_{sgmt}, \Delta\overline{\varepsilon}_{sgmt}\right) = 0 \quad \text{and} \quad Cov\left(\Delta\overline{\hat{\mu}}_{sgmt}, \Delta\overline{\varepsilon}^{\mu}_{sgmt}\right) = 0. \tag{11}$$

This assumption requires that the change in mean VA within a school-grade cell is uncorrelated with the change in the average quality of students, as measured by unobserved determinants of scores and earnings. This assumption could potentially be violated by endogenous student or teacher sorting. Student sorting at an annual frequency is minimal because of the costs of changing schools. During the period we study, most students would have to move to a different neighborhood to switch schools, which families would be unlikely to do simply because a single teacher leaves or enters a given grade. While endogenous teacher sorting is plausible over long horizons, the sharp changes we analyze are likely driven by idiosyncratic shocks such as changes in staffing needs, maternity leaves, or the relocation of a spouses. Hence, we believe that (11) is a plausible restriction at high frequencies in our data and we present evidence supporting this assumption below.

Our approach complements recent work analyzing the impacts of teacher turnover on student achievement, but is the first to use turnover to validate VA models directly. Rivkin, Hanushek, and Kain (2005) identify the variance of teacher effects from differences in variances of test score gains across schools with low vs. high teacher turnover. In contrast, we identify the impacts of teachers from first moments – the relationship between changes in mean scores across cohorts and mean teacher quality – rather than second moments. Our approach does not rely on comparisons across schools with different levels of teacher turnover, which may also differ in other unobserved dimensions that could impact earnings directly. For instance, Ronfeldt et al. (2011) show that higher rates of teacher turnover lead to lower student achievement, although they do not assess whether the mean value-added of the teaching staff predicts student achievement across cohorts.[42] Jackson and Bruegmann (2009) document peer effects by analyzing whether the VA of teachers who enter or exit affects the test scores of *other* teachers' students in their school-grade cell, but do not compare changes in mean test scores by cohort to the predictions of VA models.[43]

[42] This is less of a concern in Rivkin, Hanushek, and Kain's analysis of test score impacts because they are able to test whether the variance of test score gains is higher in grades with high turnover, thereby netting out school fixed effects. This is infeasible with outcomes in adulthood, which are observed only after schooling is complete. Rivkin, Hanushek, and Kain are unable to implement the teacher switcher design we develop here because they do not have class assignment data and thus cannot estimate each teacher's individual effect μ_j, which is necessary to construct the school-grade-cohort level mean of teacher quality.

[43] The peer effects documented by Jackson and Bruegmann could in principle affect our validation of VA using

Event Studies. We begin our analysis of teaching staff changes with event studies of scores around the entry and exit of high and low VA teachers (Figure 3). Let year 0 denote the school year that a teacher enters or exits a school-grade-subject cell and define all other school years relative to that year (e.g., if the teacher enters in 1995, year 1992 is -3 and year 1997 is +2). We define an entry event as the arrival of a teacher who did not teach in that school-grade-subject cell for the three preceding years; analogously, we define an exit event as the departure of a teacher who does not return to the same school-grade-subject cell for at least three years. We estimate VA for each teacher using only data *outside* the six-year window used for the event studies to eliminate bias due to correlated estimation errors.[44] We define a teacher as "high VA" if her estimated VA based on years outside the event study window is in the top 5% of the distribution for her subject; a "low VA" teacher has an estimated VA in the bottom 5%.[45] To obtain a balanced sample, we analyze events for which we have data on average test scores at the school-grade-subject level for at least three years before and three years after the event.[46] Because these balanced event studies require data over several years, we use the full school district data spanning 1991-2009 (rather than only the analysis sample linked to the tax data), excluding school-grade-subject cells in which we have no information on teachers.

Figure 3a plots the impact of the entry of a high-VA teacher on mean test scores. The solid series plots school-grade-subject-year means of test scores in the three years before and after a high-VA teacher enters the school-grade-subject cell, with year fixed effects removed to eliminate any secular trends.[47] We do not condition on any other covariates in this figure: each point simply shows average test scores for different cohorts of students within a school-grade-subject cell adjusted for year effects. When a high-VA teacher arrives, end-of-year test scores in the subject and grade

the switcher design. However, peer learning effects are likely to be smaller with teacher exits than entry, provided that knowledge does not deteriorate very rapidly. We find that teacher entry and exit yield broadly similar results, suggesting that spillovers across teachers are not a first-order source of bias for our technique.

[44] More precisely, we calculate VA for each teacher in each year excluding a five year window (two years prior, the current year, and two years post). Coupled with our definitions of entry and exit – which require that the teacher not be present in the school-grade-subject cell for 3 years before or after the event – this ensures that we do not use any data from the relevant cell between event years -3 and +2 to compute teacher VA.

[45] In cases where multiple teachers enter or exit at the same time, we use the teachers' mean VA in decided whether it falls in the top or bottom 5% of the VA distribution. To eliminate potential selection bias, we include high VA outliers in these event studies and our cross-cohort research design more generally; that is, we do not drop the top 2% outliers who may achieve test score gains via manipulation as we do in our baseline analysis that exploits variation across classrooms. Excluding these outliers yields very similar conclusions, as can be seen from Figure 4, which shows that changes in VA predict changes in test scores accurately throughout the value-added distribution.

[46] In school-grade-subject cells with multiple events (e.g. entry of a high VA teacher in both 1995 and 1999), we include all such events by stacking the data and using the three years before and after each event.

[47] We remove year fixed effects in this and all other event study graphs by regressing mean test scores on year dummies, computing residuals, and adding back the mean test score in the estimation sample to facilitate interpretation of the scale.

taught by that teacher rise immediately. The null hypothesis that test scores do not change from year -1 to year 0 is rejected with $p < 0.001$, with standard errors clustered by school-cohort as above. The magnitude of the increase in test scores, which is 0.036 SD from year -1 to year 0, is very similar to what one would forecast based on the change in mean teacher VA. Mean VA rises by 0.044 SD from year -1 to year 0.[48] The estimate in Column 1 of Table 2 based on cross-classroom variation implies that we should expect this increase in teacher VA to increase students' scores by $0.044 \times 0.861 = 0.038$ SD.[49] The hypothesis that the observed change in mean scores of 0.036 equals the predicted change of 0.038 is not rejected ($p = 0.76$).

Figure 3a implies that value-added accurately measures teachers' impacts on students' test scores under the identification assumption in (11). We evaluate this assumption by examining test scores for the same cohort of students in the previous school year. For example, the entry of a high-VA teacher in grade 5 in 1995 should have no impact on the same cohort's 4th grade test scores in 1994. The dashed line in Figure 3a plots test scores in the previous grade for the same cohorts of students. Test scores in the prior grade remain stable across cohorts both before and after the new teacher arrives, supporting our view that school quality and student attributes are not changing sharply around the entry of a high-VA teacher.[50]

The remaining panels of Figure 3 repeat the event study in Panel A for other types of arrivals and departures. Figure 3b examines current and lagged test scores around the departure of a high-VA teacher. There is a smooth negative trend in both current and lagged scores, suggesting that high-VA teachers leave schools that are declining in quality. However, scores in the grade taught by the teacher drop sharply relative to prior scores in the event year, showing that the departure of the high quality teacher lowers the achievement of subsequent cohorts of students. Figures 3c and 3d analyze the arrival and departure of low VA teachers. Test scores in the grade taught by the teacher fall sharply relative to prior-year scores when low VA teachers enter a school-grade cell and rise sharply when low VA teachers leave. In every case, the magnitude of the test score change is significantly different from 0 with $p < 0.001$ but is not significantly different from what one would

[48] When computing this change in mean VA, we weight teachers by the number of students they teach. For teachers who do not have any VA measures from classrooms outside the leave-out window, we impute VA as the mean leave-out VA in the sample. For a small fraction of students for whom we have no teacher information (5% of observations), we also impute teacher VA as the sample mean.

[49] We expect the observed change in scores when a high VA teachers enters to be smaller than the change in mean VA for the same reason that the cross-class coefficient is less than 1 – namely that teacher VA likely changes over time, and we use data from at least three years before or after the event to estimate teacher VA. Hence, the appropriate test for bias is whether the change in test scores matches what one would predict based on the cross-class coefficient of 0.861.

[50] We also find that class size does not change significantly around the entry and exit events we study.

forecast based on the change in mean teacher VA.[51] Together, these event studies provide direct evidence that deselecting low VA teachers and retaining high-VA teachers improves the academic achievement of students.

Teaching Staff Changes. The event studies focus on the tails of the teacher VA distribution and thus exploit only a small fraction of the variation arising from teacher turnover in the data. We now exploit all the variation due to teaching staff changes to obtain a broader estimate of the degree of bias in VA measures. To do so, we first estimate VA for each teacher using data excluding a given pair of adjacent years, $t-1$ and t. We then calculate the change in mean teacher VA for each school-grade-subject-year cell and define $\Delta\overline{\mu}_{sgmt}$ as mean teacher VA in year t minus mean teacher VA in year $t-1$. With this definition, the variation in $\Delta\overline{\mu}_{sgmt}$ is driven purely by changes in the teaching staff and not by changes in the estimated VA for the teachers. This leave-out technique again ensures that changes in mean test scores across cohorts t and $t-1$, which we denote by $\Delta\overline{A}_{sgmt}$, are not spuriously correlated with estimation error in $\Delta\overline{\mu}_{sgmt}$.

Figure 4a plots the changes in mean test scores across cohorts $\Delta\overline{A}_{sgmt}$ against changes in mean teacher value-added $\Delta\overline{\mu}_{sgmt}$. As in the event studies, we remove year fixed effects so that the estimate is identified purely from differential changes in teacher quality across school-grade-subject cells over time. For comparability with the estimates in Table 2, we only use data from the linked analysis sample in this figure. Changes in the quality of the teaching staff strongly predict changes in test scores across consecutive cohorts of students in a school-grade-subject cell. The estimated coefficient on $\Delta\overline{\mu}_{sgmt}$ is 0.843, with a standard error of 0.053 (Table 4, Column 1). This estimate is very similar to the coefficient of 0.861 obtained from the cross-class out-of-sample forecast in Column 1 of Table 2. The point estimate of the degree of bias is 2% and is not statistically distinguishable from 0. At the lower bound of the 95% confidence interval, we reject bias of more than 14%.

Figures 4b through 4d evaluate the identification assumption in (11) underlying our research design using additional placebo tests. Each of these panels replicates Figure 4a with a different dependent variable; the corresponding regression estimates are reported in Columns 2-4 of Table 4. Figure 4b shows that changes in the quality of the teaching staff are unrelated to changes in parent characteristics, as captured by the predicted score measure used in Column 2 of Table 2. In Figures 4c and 4d, we examine the impact of changes in the teaching staff in one subject on mean scores

[51] The event studies in Figure 3 pool variation from teachers switching within schools, across schools, and out of the district. Teacher switches across grades within schools have similar impacts on test scores to teacher switches out of schools.

in the *other* subject. Here, it is important to distinguish between elementary and middle schools. In elementary school, students have one teacher for both math and English. Because elementary school teachers' math and English VA are highly correlated ($r = 0.59$), changes in mean teacher VA across cohorts are highly correlated across the two subjects. But students have different teachers for the two subjects in middle school, and changes in mean VA across cohorts in one subject are thus uncorrelated with changes in mean VA in the other subject. Hence, if (11) holds, we would expect changes in mean teacher VA in English to have much smaller effects on test scores in math (and vice versa) in middle school relative to elementary school. Figures 4c and 4d show that this is indeed the case. In elementary school, changes in mean teacher VA across cohorts strongly predict changes in test scores in the other subject ($t = 11.9$, $p < 0.001$), whereas in middle schools, the coefficient is near zero and statistically insignificant ($t = 0.04$, $p = 0.97$).

Given the results of these placebo tests, any violation of (11) would have to be driven by selection on unobserved determinants of test scores that have no effect on prior test scores and only affect the subject in which teaching staff changes occur. We believe that such selection is implausible given the information available to teachers and students and the constraints they face in sorting across schools at high frequencies.

Finally, we use our quasi-experimental design to evaluate how the choice of controls affects the degree of bias in VA estimates. The results of this analysis are reported in the last column of Table 3. For comparability, we estimate the models on a constant sample of observations for which the covariates required to estimate all the models are available. Row 1 recalculates the degree of bias – defined as the percentage difference between the cross-cohort and cross-class VA coefficients as above – on this sample for the baseline model. Rows 2 and 3 show that the degree of bias is very similar when parental controls and twice-lagged test scores are including in the control vector, consistent with the very high correlations between these VA estimates and the baseline estimates discussed above. In row 4, we include only the controls that are a function of prior-year test scores: cubic polynomials in student, classroom, and school-grade math and English scores interacted with grade level. These VA estimates remain fairly highly correlated with the baseline estimates but have a somewhat larger degree of bias (14%). Finally, row 5 estimates VA without any controls at all, i.e. using raw mean test scores by teacher. These VA estimates are very poorly correlated with the other VA measures and are biased by nearly 90%. We conclude that most of the bias in VA estimates is eliminated by controlling for lagged test scores, and that further controls for demographic variables typically available in school district datasets bring the bias close to zero.

4.5 Relationship to Prior Work

Our results on the validity of VA measures reconcile the conflicting findings of prior work, including Kane and Staiger (2008) and Rothstein (2010). Rothstein reports two important results, both of which we replicate in our data. First, there is significant grouping of students into classrooms based on twice-lagged scores (lagged gains), even conditional on once-lagged scores (Rothstein 2010, Table 4). Second, this grouping on lagged gains generates minimal bias in VA estimates: controlling for twice-lagged scores does not have a significant effect on VA estimates (Rothstein 2010, Table 6; Kane and Staiger 2008, Table 6).[52] The results from our tests in Table 2 and Figure 2 are consistent with Rothstein's conclusions. Therefore, the literature is in agreement that VA measures do not suffer from bias due to selection on observables.

Rothstein quite appropriately emphasizes that his findings raise serious concerns about the *potential* for bias due to selection on unobservable student characteristics.[53] Kane and Staiger's point estimates from a randomized experiment suggest that selection on unobservables is relatively small. Our quasi-experimental tests based on teaching staff changes confirm that the bias due to selection on unobservables turns out to be negligible with greater precision. In future work, it may be useful to explore why the grouping on lagged gains documented by Rothstein is not associated with significant selection on unobservables in practice. However, the findings in this paper and prior work are sufficient to conclude that standard estimates of teacher VA can provide accurate forecasts of teachers' average impacts on students' test scores.

Note that our test, like the experiment implemented by Kane and Staiger, evaluates the accuracy of VA measures on average across teachers. It is conceivable that VA measures are biased against some subgroups of teachers and that this bias is offset by a second source of bias which is negatively correlated with true value-added (Rothstein 2009, page 567). We focus on the accuracy of average forecasts in this paper because our analysis of long-term impacts primarily evaluates the mean impacts of teacher value-added on students. A fruitful direction for future work would be to adapt the methods we propose here to evaluate the accuracy and predictive content of VA measures for

[52] An interesting question is how Rothstein's two findings are consistent with each other. There are two explanations for this pattern. First, the degree of grouping that Rothstein finds on $A_{ig,t-2}$ has small effects on residual test score gains because the correlation between $A_{ig,t-2}$ and A_{igt} conditional on $A_{ig,t-1}$ is relatively small. Second, if the component of $A_{ig,t-2}$ on which there is grouping is not the same as the component that is correlated with $A_{i,t}$, VA estimates may be completely unaffected by grouping on $A_{i,t-2}$. For both reasons, one cannot infer from grouping on $A_{i,t-2}$ that VA estimates are significantly biased by selection on $A_{i,t-2}$. See Goldhaber and Chaplin (2012) for further discussion of these and related issues.

[53] To be clear, this was the original lesson from Rothstein (2010). In personal correspondence, Rothstein notes that his findings are "neither necessary nor sufficient for there to be bias in a VA estimate" and that "if the selection is just on observables, the bias is too small to matter. The worrying scenario is selection on unobservables."

subgroups of the population.

5 Impacts of Value-Added on Outcomes in Adulthood

The results in the previous section show that value-added is a good proxy for a teacher's ability to raise students' test scores. In this section, we analyze whether value-added is also a good proxy for teachers' long run impacts. We do so by regressing outcomes in adulthood Y_i on teacher quality $\hat{\mu}_{j(i,g)}$ and observable characteristics, as in (9). We begin by pooling the data across all grade levels and then present results that estimate grade-specific coefficients on teacher VA. Recall that each student appears in our dataset once for every subject-year with the same level of Y_i but different values of $\mu_{j(i,g)}$. Hence, in this pooled regression, the coefficient estimate β represents the mean impact of having a higher VA teacher for a *single* grade between grades 4-8. We account for the repeated student-level observations by clustering standard errors at the school-cohort level as above.

We first report estimates based on comparisons of students assigned to different teachers, which identifies the causal impact of teachers under Assumption 2. We then evaluate this identification assumption by comparing these estimates to those obtained from the teacher switcher research design, which isolates quasi-experimental variation in teacher VA. We analyze impacts of teacher VA on three sets of outcomes: college attendance, earnings, and other indicators such as teenage birth rates.

5.1 College Attendance

We begin by analyzing the impact of teacher VA on college attendance at age 20, the age at which college attendance rates are maximized in our sample. In all figures and tables in this section, we condition on the standard classroom-level controls as in Figure 1.

Figure 5a plots college attendance rates at age 20 against teacher VA. Being assigned to a higher VA teacher in a single grade raises a student's probability of attending college significantly. The null hypothesis that teacher VA has no effect on college attendance is rejected with a t-statistic above 7 ($p < 0.001$). To interpret the magnitude of the impact, recall that a 1 SD increase in teacher VA raises students' test scores by 0.1 SD on average across math and English. Because we measure teacher quality μ_j in units of student test scores, a 1 unit increase in μ_j corresponds to a 10 SD increase in teacher VA. Hence, dividing the regression coefficients β by 10 yields a rough estimate of the impact of a 1 SD increase in teacher VA on the outcome of interest. In the

case of college attendance, $\beta = 4.92\%$, implying that a 1 SD better teacher in a single grade raises the probability of being in college by 0.49% at age 20, relative to a mean of 37.8%. This impact of a 1.25% increase in college attendance rates for a 1 SD better teacher is roughly similar to the impacts on other outcomes we document below.

To confirm that the relationship in Figure 5a reflects the causal impact of teachers rather than selection bias, we implement tests analogous to those in the previous section in Table 5. As a reference, the first column replicates the OLS regression estimate reported in Figure 5a. In column 2, we replace actual college attendance with predicted attendance based on parent characteristics, constructed in the same way as predicted scores above. The estimates show that one would not have predicted any significant difference in college attendance rates across students with high vs. low VA teachers based on parent characteristics.

To account for potential bias due to unobservables, we exploit quasi-experimental variation from changes in teaching staff as above. Column 3 regresses changes in mean college attendance rates across adjacent cohorts within a school-grade-subject cell on the change in mean teacher VA due to teacher staff changes $\Delta\overline{\mu}_{sgmt}$, defined as in Table 4. As above, we include no controls other than year effects. Students who happen to be in a cohort in their school that is taught by higher VA teachers are significantly more likely to go to college. The estimate of $\beta = 6.1\%$ from this quasi-experimental variation is similar to that obtained from the cross-classroom comparison in column 1, though less precise because it exploits much less variation. The null hypothesis that $\beta = 0$ is rejected with $p < 0.01$, while the hypothesis that β is the same in columns 1 and 3 is not rejected. This finding provides further evidence that teacher VA has a causal impact on college attendance rates and confirms that comparisons across classrooms with high and low VA teachers yield consistent estimates of teachers' impacts.[54]

Next, we analyze whether high-VA teachers also improve the quality of colleges that their students attend. We quantify college quality using the age 30 earnings of students who previously attended the same college, as described in Section 3. Students who do not attend college are assigned the mean earnings of individuals who do not attend college. Figure 5b plots this earnings-based index of college quality (based on the colleges students attend at age 20) vs. teacher VA. Again, there is a highly significant relationship between the quality of colleges students attend and

[54] This result rules out bias due to omitted variables that affect long-term outcomes but not test scores. For instance, one may be concerned that students who are assigned to better teachers in one subject are also assigned to better teachers in other subjects or better extracurricular activities, which would inflate estimates of long-term impacts. The cross-cohort research design rules out such biases because fluctuations in teaching staff are highly subject-specific and are uncorrelated with other determinants of student outcomes, as shown in Figure 4d.

the quality of the teachers they had in grades 4-8 ($t = 9.5$, $p < 0.001$). A 1 SD improvement in teacher VA (i.e., an increase of 0.1 in μ_j) raises college quality by \$164 (0.66%) on average (Column 4 of Table 5). Column 5 shows that exploiting the cross-cohort teacher switcher variation again yields similar estimates of the impact of teacher VA on college quality.

The \$164 estimate combines intensive and extensive margin responses because it includes the effect of increased college attendance rates on projected earnings. Isolating intensive margin responses is more complicated because of selection bias: students who are induced to go to college by a high-VA teacher will tend to attend lower-quality colleges, pulling down mean earnings conditional on attendance. We take two approaches to overcome this selection problem and identify intensive-margin effects. First, we define an indicator for "high quality" colleges as those with average earnings above the median among colleges that students attend in our sample, which is \$39,972. We regress this indicator on teacher VA in the full sample, including students who do not attend college. Column 6 of Table 5 shows that high-VA teachers increase the probability that students attend high quality colleges. A 1 SD increase in teacher VA raises the probability of attending a high quality college by 0.36%, relative to a mean of 17%. This increase is most consistent with an intensive margin effect, as students would be unlikely to jump from not going to college at all to attending a high quality college. Second, we derive a lower bound on the intensive margin effect by assuming that those who are induced to attend college attend a college of average quality. The mean college quality conditional on attending college is \$38,623, while the quality for all those who do not attend college is \$16,361. Hence, at most $(38,623 - 16,361) \times 0.49\% = \$109$ of the \$164 impact is due to the extensive margin response, confirming that teachers improve the quality of colleges that students attend.

Figure 5c shows the impact of teachers on college attendance at other ages. Teacher VA has a significant impact on the college attendance rate through age 25, partly reflecting attendance of graduate or professional schools. The impacts on college attendance at age 25 are smaller in magnitude (0.28% per 1 SD of teacher VA) than at age 20 because the mean college attendance rate at age 25 is 18.1% in this sample (Column 7 of Table 5). These continued impacts on college attendance in the mid 20's affect our analysis of earnings impacts, to which we now turn.

5.2 Earnings

The correlation between annual earnings and lifetime income rises rapidly as individuals enter the labor market and begins to stabilize only in the late twenties. We therefore begin by analyzing the

impacts of teacher VA on earnings at age 28, the oldest age at which we have a sufficiently large sample of students to obtain precise estimates.[55] Figure 6 plots earnings at age 28 against teacher VA, conditioning on the same set of classroom-level controls as above. Being assigned to a higher value-added teacher has a clear, statistically significant impact on earnings, with the null hypothesis of $\beta = 0$ rejected with $p < 0.01$. A 1 SD increase in teacher VA in a single grade increases earnings at age 28 by \$182, 0.9% of mean earnings in the regression sample. This regression estimate is also reported in Column 1 of Table 6. Column 2 shows the effect on wages at age 30. The point estimate is slightly larger than that at age 28, but because the sample is only one-sixth the size, the 95% confidence interval for the estimate is very wide. We therefore focus on earnings impacts up to age 28 for the remainder of our analysis.

To interpret the magnitude of the effect of teacher VA on earnings at age 28, we calculate the lifetime earnings impact of having a 1 SD higher VA teacher in a single grade. We assume that the percentage gain in earnings remains constant at 0.9% over the life-cycle and that earnings are discounted at a 3% real rate (i.e., a 5% discount rate with 2% wage growth) back to age 12, the mean age in our sample. Under these assumptions, the mean present value of lifetime earnings at age 12 in the U.S. population is approximately \$522,000.[56] Hence, the financial value of having a 1 SD higher VA teacher (i.e., a teacher at the 84th percentile instead of the median) is $0.9\% \times \$522,000 \simeq \$4,600$ per grade.[57] Another useful benchmark is the increase in earnings from an additional year of schooling, which is around 6% per year (see e.g., Oreopoulos 2006). Having a teacher in the first percentile of the value-added distribution (2.33 SD below the mean) for one year thus has an earnings impact equivalent to attending school for about 60% of the school year. This magnitude is plausible, insofar as attending school even with very low quality teaching is likely to have some returns due to benefits from peer interaction and other factors.

Next, we analyze how teacher value-added affects the trajectory of earnings by examining earnings impacts at each age from 20 to 28. We run separate regressions of earnings at each age on teacher VA and the standard vector of classroom controls. Figure 7a plots the coefficients from these regressions (which are reported in Appendix Table 10), divided by average earnings at each

[55] Although individuals' earnings trajectories remain quite steep at age 28, earnings levels at age 28 are highly correlated with earnings at later ages (Haider and Solon 2006), a finding we confirm in the tax data (Chetty et al. 2011, Appendix Table I).

[56] We calculate this number using the mean wage earnings of a random sample of the U.S. population in 2007 to obtain an earnings profile over the lifecycle, and then inflate these values to 2010 dollars (see Chetty et al. 2011 for details).

[57] The undiscounted earnings gains (assuming a 2% growth rate but 0% discount rate) are approximately \$25,000 per student.

age to obtain percentage impacts. As above, we multiply the estimates by 0.1 to interpret the effects as the impact of a 1 SD increase in teacher VA. The impact of teacher quality on earnings rises almost monotonically with age. At early ages, the impact of higher VA is *negative* and significant, which is consistent with our finding that higher VA teachers induce their students to go to college. As these students enter the labor force, they have steeper earnings trajectories and eventually earn significantly more than students who had lower VA teachers in grades 4-8. The earnings impacts become positive and statistically significant starting at age 26. By age 28, the earnings impact is nearly 1% of earnings, as in Figure 7. Stated differently, higher teacher VA increases the growth rate of earnings when students are in their 20s. In column 3 of Table 6, we verify this result by regressing the change in earnings from age 22 to age 28 on teacher VA. As expected, a 1 SD increase in teacher VA increases earnings growth by $180 (1.3%) over this period.

We obtain further insight into the role of college in mediating these changes in earnings trajectories by comparing the impacts of teacher VA on students who attend grade schools with low vs. high college attendance rates. We divide the sample into two groups: students who attend schools with an age 20 college attendance rate above vs. below 35%, the sample mean. In schools with low college attendance rates at age 20, few students are in college at age 25. As a result, teacher VA does not have a significant impact on college attendance rates at age 25 for students in these schools, as shown in Column 4 of Table 6. In contrast, in schools with high college attendance rates, a 1 SD increase in teacher VA raises college attendance rates by 0.47 percentage points even at age 25. If college attendance masks earnings impacts, we should expect the effects of teacher VA on wage growth to be higher in these high college attendance schools.

Figure 7b tests this hypothesis by plotting the effect of value-added on earnings by age for students who attended schools with above- and below- average college attendance rates. As expected, the impacts of teacher VA on earnings rise much more sharply with age for students who attended grade schools with high college attendance rates. Teacher VA has a negative impact on earnings in the early 20's for students who attended such schools, whereas its impacts are always positive for students who attended schools with low college attendance rates. The positive impacts of teacher VA on earnings even in subgroups that are unlikely to attend college indicates that better teaching has direct returns in the labor market independent of its effects on college attendance. Columns 6 and 7 of Table 6 confirm that the effect of teacher VA on wage growth from age 22 to 28 is much larger for students who attended schools with high college attendance rates.

The results in Figure 7 suggest that the 0.9% mean earnings impact per SD of teacher VA

at age 28 may understate the impact on lifetime earnings, particularly for high SES groups. To gauge how much further the earnings impacts might rise over time, we use the cross-sectional correlation between test scores and earnings, which we can estimate with greater precision up to age 30. Appendix Table 4 lists coefficients from OLS regressions of earnings at each age on test scores. These regressions pool all grades, control for the same variables used to estimate the baseline value-added model, and use a constant sample of students for whom we observe earnings from 20-30 to eliminate cohort effects. The correlation between test scores and earnings is roughly 20% higher at age 30 than at age 28. If the causal impacts of teacher VA match these cross-sectional patterns by age, the lifetime earnings impact of a 1 SD improvement in teacher VA in a single grade would likely exceed 1.1%.

The cross-sectional relationship between test scores and earnings reported in Appendix Table 4 implies that a 0.1 SD increase in test scores is associated with a 1.1% increase in earnings at age 28. Hence, the impact of teacher VA is similar to the impact one would have predicted based on the impact of VA on end-of-grade test scores and the cross-sectional relationship between test scores and earnings. This result aligns with previous evidence that improvements in education raise contemporaneous scores, then fade out in later scores (as shown in Figure 2), only to reemerge in adulthood (Deming 2009, Heckman et al. 2010c, Chetty et al. 2011).

5.3 Other Outcomes

We now analyze the impacts of teacher VA on other outcomes, starting with our "teenage birth" measure, which is an indicator for filing a tax return and claiming a dependent who was born while the mother was a teenager (see Section 3). We first evaluate the cross-sectional correlations between this proxy for teenage birth and test scores as a benchmark. Students with a 1 SD higher test score are 3.8 percentage points less likely to have a teenage birth relative to a mean of 8% (Appendix Table 3). Conditional on lagged test scores and other controls, a 1 SD increase in test score is associated with a 1 percentage point reduction in teenage birth rates. These correlations are significantly larger for populations that have a higher risk of teenage birth, such as minorities and low-income students (Appendix Table 5). These cross-sectional patterns support the use of this measure as a proxy for teenage births even though we can only identify children who are claimed as dependents in the tax data.

Column 1 of Table 7 analyzes the impact of teacher VA on the fraction of female students who have a teenage birth. Having a 1 SD higher VA teacher in a single year from grades 4 to 8 reduces

the probability of a teen birth by 0.099 percentage points, a reduction of roughly 1.25%, as shown in Figure 8a. This impact is very similar to the cross-sectional correlation between scores and teenage births, echoing our results on earnings and college attendance.

Column 2 of Table 7 analyzes the impact of teacher VA on the socio-economic status of the neighborhood in which students live at age 25, measured by the percent of college graduates living in that neighborhood. A 1 SD increase in teacher VA raises neighborhood SES by 0.063 percentage points (0.5% of the mean) by this metric, as shown in Figure 8b. Column 3 shows that this impact on neighborhood quality more than doubles at age 28, consistent with the growing earnings impacts documented above.

Finally, we analyze impacts on retirement savings. Teacher VA does not have a significant impact on 401(k) savings at age 25 in the pooled sample (not reported). However, Column 4 shows that for students who attended schools with low college attendance rates (defined as in Column 4 of Table 6), a 1 SD increase in teacher VA raises the probability of having a 401(k) at age 25 by 0.19 percentage points (1.6% of the mean). In contrast, Column 5 shows that for students in high college-attendance schools, the point estimate of the impact is negative. These results are consistent with the impacts on earnings trajectories documented above. In schools with low college attendance rates, students who get high-VA teachers find better jobs by age 25 and are more likely to start saving in 401(k)'s. In schools with high college attendance rates, students who get high-VA teachers are more likely to be in college at age 25 and thus may not obtain a job in which they begin saving for retirement until they are older.

5.4 Heterogeneity Analysis

In Table 8, we analyze whether teacher value-added has heterogeneous effects across demographic groups and subjects. We study impacts on college quality at age 20 rather than earnings because the heterogeneity analysis requires large samples and because the college quality measure provides a quantitative metric based on projected earnings gains.

Panel A studies impact heterogeneity across population subgroups. Each number in the first row of the table is a coefficient estimate from a separate regression of college quality on teacher VA, with the same classroom-level controls as in the previous sections. Columns 1 and 2 consider heterogeneity by gender. Columns 3 and 4 consider heterogeneity by parental income, dividing students into groups above and below the median level of parent income in the sample. Columns 5 and 6 split the sample into minority and non-minority students.

Two lessons emerge from Panel A of Table 8. First, the point estimates of the impacts of teacher VA are larger for girls than boys, although one can reject equality of the impacts only at a 10% significance level. Second, the impacts are larger for higher-income and non-minority households in absolute terms. For instance, a 1 SD increase in VA raises college quality by $123 for children whose parents have below-median income, compared with $209 for those whose parents have above-median income. However, the impacts are much more similar as a percentage of mean college quality: 0.56% for low-income students vs. 0.75% for high-income students.

The larger dollar impact for high socioeconomic students could be driven by two channels: a given increase in teacher VA could have larger impacts on the test scores of high SES students or a given increase in scores could have larger long-term impacts. The second row of coefficient estimates of Table 8 shows that the impacts of teacher VA on scores are virtually identical across all the subgroups in the data. In contrast, the correlation between scores and college quality is significantly larger for higher SES students (Appendix Table 5). Although not conclusive, these findings suggest that the heterogeneity in teachers' long term impacts is driven by the second mechanism, namely that high SES students benefit more from test score gains. Overall, the heterogeneity in treatment effects indicates that teacher quality is complementary to family inputs and resources, i.e. the marginal value of better teaching is *larger* for students from high SES families. An interesting implication of this result is that higher income families should be willing to pay more for teacher quality.

Panel B of Table 8 analyzes differences in teachers' impacts across subjects. For these regressions, we split the sample into elementary (Columns 1-3) and middle (Columns 4-6) schools. We first analyze the effects of teacher VA in each subject separately on a constant sample with a fixed set of controls and then include both math and English teacher VA in the same regression. In all the specifications, the coefficients on VA are larger in English than math. An English teacher who raises her students' test scores by 1 SD has a larger long-term impact than a math teacher who generates a commensurate test score gain. However, it is important to recall that the variance of teacher effects is larger in math than English: a 1 SD improvement in teacher VA raises math test scores by approximately 0.118 SD, compared with 0.081 SD in English. Hence, a 1 SD increase in the quality of a math teacher actually has a relatively similar impact to a 1 SD increase in the quality of an English teacher.

Including both English and math VA in the same regression has very different effects in elementary vs. middle school. As discussed above, students have one teacher for both subjects in

elementary school but not middle school. Because a given teacher's math and English VA are highly correlated ($r = 0.59$), the magnitude of the two subject-specific coefficients drops by nearly 40% when included together in a single regression for elementary school (Column 3). Intuitively, when math VA is included by itself in elementary school, it partly picks up the effect of having better teaching in English as well. In contrast, including both math and English teacher VA in middle school has a much smaller effect on the estimates, as shown in Column 6.

5.5 Robustness Checks

We conclude our empirical analysis by assessing the robustness of our results to alternative empirical specifications, focusing on the simplifications we made for computational tractability.

First, we assess the robustness of our statistical inferences to alternative forms of clustering standard errors. Appendix Table 7 reports alternative standard error calculations for three of our main specifications: the impact of teacher VA on scores, college attendance at age 20, and earnings at age 28. We estimate each of these models using the baseline control vector used in Table 2. Panels A of Appendix Table 7 shows that a block bootstrap at the student level, which accounts for repeated student observations, yields narrower confidence intervals than school-cohort clustering. Panel B shows that in smaller subsamples of our data, two-way clustering by class and student yields slightly smaller standard errors than school-cohort clustering. Panel C shows that school-cohort clustering is also conservative relative to clustering by classroom in a sample that includes only the first observation for each student.

Second, we assess the robustness of our estimates to alternative control vectors (Panel D of Appendix Table 7). Including the student-level controls used when estimating the VA model in addition to the baseline classroom-level control vector used to estimate the regressions in Tables 2, 5, and 6 has virtually no impact on the coefficients or standard errors. The last row of the table evaluates the impacts of including school by year fixed effects. In this row, we include school by year effects both when estimating VA and in the second-stage regressions of VA on adult outcomes. The inclusion of school by year fixed effects does not affect our qualitative conclusion that teacher VA has substantial impacts on adult outcomes, but the estimated impact on college attendance at age 20 falls, while the impact on earnings at age 28 rises.[58]

Finally, we replicate the baseline results using raw estimates of teacher quality without the

[58] We did not include school-year fixed effects in our baseline specifications because school districts typically seek to rank teachers within their districts rather than within schools. Moreover, our tests in Section 4 suggest that such fixed effects are not necessary to obtain unbiased estimates of the impacts of teacher VA.

Empirical Bayes shrinkage correction, denoted by $\overline{\nu}_j$ in Section 2. We again exclude the current year when estimating $\overline{\nu}_j$ to account for correlated estimation error as above. In columns 1-4 of Appendix Table 11, we estimate specifications analogous to (9) using OLS, with a leave-year-out measure $\overline{\nu}_j^t$ on the right hand side instead of $\widehat{\mu}_j^t$. The estimated coefficients are roughly half of those reported above, reflecting the substantial attenuation from measurement error in teacher quality. The shrinkage correction implemented in our baseline measure of teacher VA is one approach to correct for this measurement error. As an alternative approach, we regress each outcome on test scores, instrumenting for scores using the raw teacher effects $\overline{\nu}_j$. The resulting two-stage least squares coefficients are reported in Columns 5-7 of Appendix Table 11. These 2SLS estimates are very similar to our baseline results, confirming that our findings are not sensitive to the way in which correct for measurement error in teacher quality.

6 Policy Calculations

In this section, we use our estimates to answer two policy questions. First, do teachers matter more in some grades relative to others? Second, what are the expected earnings gains from retaining or deselecting teachers based on their estimated VA?

6.1 Impacts of Teachers by Grade

The reduced-form estimates in the previous section identify the impacts of replacing a single teacher j with another teacher j' in one classroom. While this question is of interest to parents, policymakers are typically interested in the impacts of reforms that improve teacher quality more broadly. As shown in (5), the reduced-form impact of changing the teacher of a single classroom includes the impacts of being tracked to a better teacher in subsequent grades. While a parent may be interested in the reduced-form impact of teacher VA in grade g (β_g), a policy reform that raises teacher quality in grade g will not allow every child to get a better teacher in grade $g+1$. In this section, we estimate teachers' net impacts in each grade, holding fixed future teacher VA ($\tilde{\beta}_g$), to shed light on this policy question.

Because we have no data after grade 8, we can only estimate teachers' net effects holding fixed teacher quality up to grade 8.[59] We therefore set $\tilde{\beta}_8 = \beta_8$. We recover $\tilde{\beta}_g$ from estimates of β_g by subtracting out the impacts of future teachers on earnings iteratively. Consider the effect of

[59] If tracking to high school teachers is constant across all grades in elementary school, our approach accurately recovers the relative impacts of teachers in grades 4-8.

teacher quality in 7th grade. Our reduced-form estimate of β_7, obtained by estimating (9) using only grade 7, can be decomposed into two terms:

$$\beta_7 = \tilde{\beta}_7 + \rho_{78}\tilde{\beta}_8$$

where ρ_{78} is the extent to which teacher VA in grade 7 increases teacher VA in grade 8. We can estimate $\hat{\rho}_{78}$ using an OLS regression that parallels (9) with future teacher VA as the dependent variable:

$$\hat{\mu}_{j(i,8)} = \alpha + \hat{\rho}_{78}\hat{\mu}_{j(i,7)} + f_1(A_{i,t-1}) + f_2(e_{j(i,7,t)}) + \phi_1 X_{i7t} + \phi_2 \bar{X}_{c(i,7,t)} + \eta^{\mu}_{it78}.$$

Combining these two equations shows that the net impact of the grade 7 teacher is simply her reduced-form impact minus her indirect impact via tracking to a better 8th grade teacher:

$$\tilde{\beta}_7 = \beta_7 - \hat{\rho}_{78}\beta_8.$$

Iterating backwards, we can calculate $\tilde{\beta}_6$ by estimating $\hat{\rho}_{68}$ and $\hat{\rho}_{67}$ and so on until we obtain the full set of net impacts. We show formally that this procedure recovers net impacts $\tilde{\beta}_g$ in Appendix B.

This approach to calculating teachers' net impacts has three important limitations. First, it assumes that all tracking to future teachers occurs via teacher VA on test scores. We allow students who have high-VA teachers in grade g to be tracked to higher VA ($\mu_{j(i,g+1)}$) teachers in grade $g+1$, but *not* to teachers with higher unobserved earnings impacts μ^Y. We are forced to make this strong assumption because we have no way to estimate teacher impacts on earnings that are orthogonal to VA, as discussed in Section 2. Second, $\tilde{\beta}_g$ does not net out potential changes in other factors besides teachers, such as peer quality or parental inputs. Hence, $\tilde{\beta}_g$ cannot be interpreted as the "structural" impact of teacher quality holding fixed all other inputs in a general model of the education production function (e.g., Todd and Wolpin 2003). Finally, our approach assumes that teacher effects are additive across grades. We cannot identify complementarities in teacher VA across grades because our identification strategy forces us to condition on lagged test scores, which are endogenous to the prior teacher's quality. It would be valuable to relax these assumptions in future work to obtain a better understanding of how the sequence of teachers one has in school affects outcomes in adulthood.

Figure 9 displays our estimates of β_g and $\tilde{\beta}_g$, which are also reported in Appendix Table 12. We use college quality (projected earnings at age 30 based on college enrollment at age 20) as

the outcome to have sufficient precision to identify grade-specific effects. We estimate β_g using specifications analogous to Column 4 of Table 5 for each grade separately. Because the school district data system did not cover many middle schools in the early and mid 1990s, we cannot analyze the impacts of teachers in grades 6-8 for more than half the students who are in 4th grade before 1994. To obtain a more balanced sample for comparisons across grades, we restrict attention to cohorts who would have been in 4th grade during or after 1994 for this analysis.

Figure 9 has two lessons. First, the net impacts $\tilde{\beta}_g$ are close to the reduced-form impacts. This is because the tracking coefficients $\rho_{g,g'}$ are generally quite small, as shown in Appendix Table 13. Tracking is slightly larger in middle school, as one would expect, but still has a relatively small impact on $\tilde{\beta}_g$. Second, teachers' long-term impacts are large and significant in all grades. Although the estimates in each grade have relatively wide confidence intervals, there is no systematic trend in the impacts. This pattern is consistent with the cross-sectional correlations between test scores and adult outcomes, which are also relatively stable across grades (Appendix Table 6).

One issue that complicates cross-grade comparisons is that teachers spend almost the entire school day with their students in elementary school (grades 4-5 as well as 6 in some schools), but only their subject period (Math or English) in middle school (grades 7-8). If teachers' skills are correlated across subjects – as is the case with math and English value-added, which have a correlation of 0.59 for elementary school teachers – then a high-VA teacher should have a greater impact on earnings in elementary school than middle school because they spend more time with the student. The fact that high-VA math and English teachers continue to have substantial impacts even in middle school underscores our conclusion that higher quality education has substantial returns well beyond early childhood.

6.2 Impacts of Selecting Teachers on VA

In this section, we use our estimates to predict the potential earnings gains from selecting and retaining teachers on the basis of their VA. The primary objective of these calculations is to illustrate the magnitudes of teachers' impacts rather than evaluate selection as a policy to improve teacher quality.

We make three assumptions in our calculations. First, we assume that the percentage impact of a 1 unit improvement in teacher VA on earnings observed at age 28, which we denote by b, remains constant over the life-cycle. Second, we do not account for general equilibrium effects that may reduce wages if all children are better educated or for non-monetary returns to education such

as reductions in teenage birth rates (Oreopoulos and Salvanes 2010). Third, we follow Krueger (1999) and discount earnings gains at a 3% real annual rate (consistent with a 5% discount rate and 2% wage growth) back to age 12, the average age in our sample. Under this assumption, the present value of earnings at age 12 for the average individual in the U.S. population is $\$522,000$, as noted above.

We first evaluate Hanushek's (2009, 2011) proposal to replace the 5 percent of teachers with the lowest value-added with teachers of average quality. To calculate the impacts of such a policy, note that a teacher in the bottom 5% of the true VA distribution is on average 2.04 standard deviations below the mean teacher quality. Therefore, replacing a teacher in the bottom 5% with an average teacher generates a gain per student of

$$\$522,000 \times 2.04 \times b\sigma_\mu$$

where σ_μ denotes the standard deviation of teacher effects. We set $b = \$1,815/20,362 = 8.9\%$ based on the estimate in Column 1 of Table 6 and $\sigma_\mu = (0.081 + 0.118)/2$, the average of the SD of teacher effects across math and English. With these values, replacing a teacher in the bottom 5% with an average teacher generates earnings gains of $\$9,422$ per student in present value at age 12, or $\$267,000$ for a class of average size (28.3 students). The undiscounted cumulative earnings gains from deselection are 5.5 times larger than these present value gains ($52,000 per student and $1.48 million per classroom), as shown in Appendix Table 14.[60] These calculations show that improving teacher VA – whether by selection, better training, or other methods – is likely to have substantial returns for students.

The $\$267,000$ present value gain is based on selecting teachers based on their *true* VA μ_j. In practice, we only observe a noisy estimate of μ_j based on a small number of classrooms. To calculate the gains from deselecting the bottom 5% of teachers based on their *estimated* VA, note that (4) implies that $\sigma_{\hat{\mu}} = \sigma_\mu\sqrt{r(n_c)}$ where $r(n_c)$ is the reliability of VA estimates based on n_c classrooms of data. Hence, with n_c years of data, the bottom 5 percent of teachers ranked on $\hat{\mu}_j$ have a mean forecasted quality of $2.04\sigma_\mu\sqrt{r(n_c)}$. The gain from deselecting the lowest 5% of teachers based on n_c classrooms of data is thus $G(n_c) = \$267,000 \cdot \sqrt{r(n_c)}$.[61]

[60] These calculations assume that deselected teachers are replaced by teachers with the same amount of experience rather than rookies. Rookie teachers' test score impacts are 0.03 SD below those of experienced teachers, on average. However, given that the median teacher remains in our data for 6 years, the expected benefits of deselection would be reduced by less than 3% ($\frac{0.03/6}{2.04\sigma_\mu}$) from hiring inexperienced teachers to replace those deselected.

[61] This calculation accounts for estimation error due to noise but ignores drift in VA over time (except for drift due to teacher experience, which we control for in our analysis). Drift affects the calculation in two ways. First, our

Figure 10 plots $G(n_c)$ assuming a constant class size of 28.3 students; see Appendix Table 14 for the values underlying this figure. It yields three lessons. First, the gains from deselecting low quality teachers on the basis of very few years of data are much smaller than the maximum attainable gain of \$267,000 because of the noise in VA estimates. With one year of data, the gains are about half as large (\$135,000). This is because reliability with one class of students is approximately $r(1) = \frac{1}{4}$ in our data, consistent with prior work on teacher effects (Staiger and Rockoff 2010, McCaffrey et al. 2009). That is, one-quarter of the variance in the mean test score residual for a single classroom is driven by teacher quality, with the remaining variance due to classroom and student level noise. Second, the gains grow fairly rapidly with more data in the first 3 years but the marginal gains from additional information are small. With three years of data, one can achieve more than 70% of the maximum impact (\$190,000). Waiting for three more years would increase the gain by \$30,000 but has an expected cost of $3 \times \$190,000 = \$570,000$. The marginal gains from obtaining one more year of data are outweighed by the expected cost of having a low VA teacher on the staff even after the first year (Staiger and Rockoff 2010). Third, because VA estimates are noisy, there could be substantial gains from using other signals of quality to complement VA estimates, such as principal evaluations or other subjective measures based on classroom observation.

An alternative approach to improving teacher quality is to increase the retention of high-VA teachers. Retaining a teacher at the 95th percentile of the estimated VA distribution (using 3 classrooms of data) for an extra year would yield present value earnings gains of $\$522,000 \times 1.96 \times b\sigma_\mu\sqrt{r(3)} = \$182,000$. In our data, roughly 9% of teachers in their third year do not return to the school district for a fourth year.[62] Clotfelter et al. (2008) estimate that a \$1,800 bonus payment in North Carolina reduces attrition rates by 17%. Based on this estimate, a one time bonus payment of \$1,800 to high-VA teachers who return for a fourth year would increase retention rates in the next year by 1.5 percentage points and generate an average benefit of \$2,730. The expected benefit of offering a bonus to even an excellent (95th percentile) teacher is only modestly larger than the cost because for every extra teacher retained, one must pay bonuses to 60 (91/1.5) additional teachers.

estimate of b uses estimated VA from other years and thereby understates the impact of a 1 unit increase in true VA on earnings. This leads us to understate the \$267,000 gain. Second, if true VA is mean reverting, deselecting teachers based on their current VA will yield smaller gains in subsequent years, because some of the low VA teachers improve over time. An interesting direction for future research is to estimate the process that VA follows and then identify the expected gains from selecting teachers based on their true VA over various horizons.

[62] The rate of attrition bears little or no relation to VA, consistent with the findings of Boyd et al. (2009).

One important caveat to these calculations is that they assume that teacher effectiveness μ_j does not vary with classroom characteristics. Our estimates of VA only identify the component of teacher quality that is orthogonal to lagged test scores and the other characteristics that we control for to account for sorting. That is, teachers are evaluated relative to the average quality of teachers with similar students, not relative to the population. Thus, while we can predict the effects of selecting teachers among those assigned to a sub-population of similar students, we cannot predict the impacts of policies that reassign teachers to randomly selected classrooms from the population (Rubin, Stuart, and Zanutto 2004). This is a limitation in all existing value-added measures of teacher quality and could have significant implications for their use if teaching quality interacts heavily with student attributes. Lockwood and McCaffrey (2009) argue that such interactions are small relative to the overall variation in teacher VA. In addition, our estimates based on teaching staff changes suggest that VA is relatively stable as teachers switch to different grades or schools. Nevertheless, further work is needed on this issue if a policymaker is considering reassigning teachers across classrooms and seeks a global ranking of their relative quality.

7 Conclusion

This paper has presented evidence that existing value-added measures are informative about teachers' long-term impacts. However, two important issues must be resolved before one can determine whether VA should be used to evaluate teachers. First, using VA measures in high-stakes evaluations could induce responses such as teaching to the test or cheating, eroding the signal in VA measures. This question can be addressed by testing whether VA measures from a high stakes testing environment provide as good of a proxy for long-term impacts as they do in our data.[63] If not, one may need to develop metrics that are more robust to such responses, as in Barlevy and Neal (2012). Districts may also be able to use data on the persistence of test score gains to identify test manipulation, as in Jacob and Levitt (2003), and thereby develop a more robust estimate of VA. Second, one must weigh the cost of errors in personnel decisions against the mean benefits from improving teacher value-added. We quantified mean earnings gains from selecting teachers on VA but did not quantify the costs imposed on teachers or schools from the turnover generated by such policies.

[63] As we noted above, even in the low-stakes regime we study, some teachers in the upper tail of the VA distribution have test score impacts consistent with test manipulation. If such behavior becomes more prevalent when VA is actually used to evaluate teachers, the predictive content of VA as a measure of true teacher quality could be compromised.

Whether or not VA should be used as a policy tool, our results suggest that parents would place great value on having their child in the classroom of a high value-added teacher. Consider a teacher whose true VA is 1 SD above the median who is contemplating leaving a school. Each child would gain approximately $25,000 in total (undiscounted) lifetime earnings from having this teacher instead of the median teacher. With an annual discount rate of 5%, the parents of a classroom of average size should be willing to pool resources and pay this teacher approximately $130,000 ($4,600 per parent) to stay and teach their children during the next school year. Our analysis of teacher entry and exit directly confirms that retaining such a high-VA teacher would improve students' outcomes.

While these calculations show that good teachers have great value, they do not by themselves have implications for optimal teacher salaries or merit pay policies. The most important lesson of this study is that finding policies to raise the quality of teaching – whether via the use of value-added measures, changes in salary structure, or teacher training – is likely to have substantial economic and social benefits in the long run.

References

1. Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in Chicago Public High Schools." *Journal of Labor Economics* 24(1): 95-135.

2. Altonji, Joseph, Todd Elder, and Christopher Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1): 151-184.

3. Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard. 2010. "Problems with the Use of Student Test Scores to Evaluate Teachers." Economic Policy Institute Briefing Paper #278.

4. Barlevy, Gadi and Derek Neal. 2012. "Pay for Percentile." *American Economic Review* (forthcoming).

5. Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff "Who Leaves? Teacher Attrition and Student Achievement." *Economics of Education Review* (forthcoming).

6. Cameron, Colin A., Jonah B. Gelbach, and Douglas Miller. 2011. "Robust Inference with Multi-way Clustering," *Journal of Business and Economic Statistics* 29 (2): 238-249.

7. Carrell, Scott E. and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy* 118(3): 409-432.

8. Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR" *Quarterly Journal of Economics* 126(4): 1593-1660, 2011.

9. Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2011. "New Evidence on the Long-Term Impacts of Tax Credits." IRS Statistics of Income White Paper.

10. Clotfelter, Charles, Elizabeth Glennie, Helen Ladd, and Jacob Vigdor. 2008. "Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina." *Journal of Public Economics* 92: 1352-70.

11. Corcoran, Sean P. 2010. "Can Teachers be Evaluated by Their Students' Test Scores? Should they Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice." Report for the Annenberg Institute for School Reform, Education Policy for Action Series.

12. Cunha, Flavio and James J. Heckman. 2010. "Investing in our Young People." NBER Working Paper 16201.

13. Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78(3): 883–931.

14. Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3): 111-134.

15. Dynarski, Susan, Joshua M. Hyman, and Diane Whitmore Schanzenbach. 2011. "Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion." NBER Working Paper 17533.

16. Goldhaber, Dan and Duncan Chaplin, 2012. "Assessing the 'Rothstein Test': Does It Really Show Teacher Value-Added Models Are Biased?" University of Washington Working Paper.

17. Goldhaber, Dan and Michael Hansen. 2010. "Using Performance on the Job to Inform Teacher Tenure Decisions." *American Economic Review* 100(2): 250-255.

18. Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job," The Hamilton Project White Paper 2006-01.

19. Haider, Steven, and Gary Solon. 2006. "Life-cycle variation in the Association Between Current and Lifetime Earnings." *American Economic Review* 96: 1308-1320.

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

21. Hanushek, Eric A. 2009. "Teacher Deselection." in Creating a New Teaching Profession, ed. Dan Goldhaber and Jane Hannaway, 165–80. Washington, DC: Urban Institute Press.

22. Hanushek, Eric A. 2011. "The Economic Value of Higher Teacher Quality." *Economics of Education Review* 30: 466–479.

23. Hanushek Eric A., John F. Kain and Steven G. Rivkin. 2004. "Why Public Schools Lose Teachers," *Journal of Human Resources* 39(2): 326-354

24. Heckman, James J. 2002. "Policies to Foster Human Capital." *Research in Economics* 54(1): 3-56.

25. Heckman, James J., Seong H. Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam. Yavitz. 2010a. "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program." *Quantitative Economics* 1(1): 1-46.

26. Heckman, James J., Seong H. Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010b. "The Rate of the Return to the High Scope Perry Preschool Program." *Journal of Public Economics* 94: 114-128.

27. Heckman, James J., Lena Malofeeva, Rodrigo Pinto, and Peter A. Savelyev. 2010c. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," University of Chicago, unpublished.

28. Holmstrom, Bengt and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organizations* 7: 24–52.

29. Internal Revenue Service. 2010. *Document 6961: Calendar Year Projections of Information and Withholding Documents for the United States and IRS Campuses 2010-2018,* IRS Office of Research, Analysis, and Statistics, Washington, D.C.

30. Jackson, C. Kirabo. 2010. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence From Teachers." NBER Working Paper No. 15990.

31. Jackson, C. Kirabo, and Elias Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," *American Economic Journal: Applied Economics* 1(4): 85–108.

32. Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89(6): 761–796.

33. Jacob, Brian A. and Steven D. Levitt. 2003. "Rotten Apples: An Investigation Of The Prevalence And Predictors Of Teacher Cheating." *The Quarterly Journal of Economics* 118(3): 843-877.

34. Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. "The Persistence of Teacher-Induced Learning Gains," Journal of Human Resources, 45(4): 915-943.

35. Jacob, Brian A. and Jonah E. Rockoff. 2011. "Organizing Schools to Improve Student Achievement: Start Times, Grade Configurations, And Teaching Assignments" Hamilton Project Discussion Paper 2011-08.

36. Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper No. 14607.

37. Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City," *Economics of Education Review* 27: 615–631

38. Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources* 46(3): 587-613.

39. Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2): 497-532.

40. Lockwood, J.R. and Daniel F. McCaffrey. 2009. "Exploring Student-Teacher Interactions in Longitudinal Achievement Data," *Education Finance and Policy* 4(4): 439-467.

41. McCaffrey, Daniel F., Tim R. Sass, J.R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy* 4(4): 572-606.

42. Morris, Carl. 1983. "Parametric Empirical Bayes Inference: Theory and Applications" *Journal of the American Statistical Association* 78: 47-55.

43. Murnane, Richard J. 1975. *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, MA: Ballinger.

44. Neal, Derek A. and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics* 92(2): 263-283.

45. Oreopoulos, Philip. 2006. "Estimating Average and Local Average Treatment Effects of Education when Compulsory School Laws Really Matter." American Economic Review 96(1): 152-175.

46. Oreopoulos, Philip, and Kjell G. Salvanes. 2010. "Priceless: The Nonpecuniary Benefits of Schooling." *Journal of Economic Perspectives* 25(1): 159–84.

47. Rivkin, Steven. G., Eric. A. Hanushek, and John F. Kain. 2005. "Teachers, Schools and Academic Achievement." *Econometrica* 73: 417–458.

48. Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review* 94: 247-252.

49. Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2011. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." *American Economic Review*, forthcoming.

50. Rockoff, Jonah E. and Cecilia Speroni. 2011. "Subjective and Objective Evaluations of Teacher Effectiveness: Evidence from New York City," *Labour Economics* 18: 687–696

51. Ronfeldt, Matthew, Hamilton Lankford, Susanna Loeb, James Wyckoff. 2011. "How Teacher Turnover Harms Student Achievement," NBER Working Paper No. 17176.

52. Rothstein, Jesse. 2009. "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy* 4(4), 537-571.

53. Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125(1): 175-214.

54. Rubin, Donald B., Elizabeth A. Stuart and Elaine L. Zanutto. 2004. "A Potential Outcomes View of Value-Added Assessment in Education." *Journal of Educational and Behavioral Statistics*, 29(1): 103-116.

55. Springer, Matthew G., Ballou, Dale, Hamilton, Laura, Le, Vi-Nhuan, Lockwood, J.R., McCaffrey, Daniel F., Pepper, Matthew, and Brian M. Stecher. 2010. "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching." Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

56. Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24: 97-117.

57. Todd, Petra E. and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement" *The Economic Journal* 113(485): F3-F33.

58. U.S. Census Bureau. 2006-2008. "American Community Survey."
ACS 3-year data. http://www.census.gov.

59. U.S. Census Bureau. 2010. "School Enrollment–Social and Economic Characteristics of Students: October 2008, Detailed." Washington, D.C.
http://www.census.gov/population/www/socdemo/school.html.

Appendix A: Matching Algorithm

We follow the matching algorithm developed in Chetty et al. (2011) to link the school district data to tax records. The algorithm was designed to match as many records as possible using variables that are not contingent on ex post outcomes. Date of birth, gender, and last name in the tax data are populated by the Social Security Administration using information that is not contingent on ex post outcomes. First name and ZIP code in tax data are contingent on observing some ex post outcome. First name data derive from information returns, which are typically generated after an adult outcome like employment (W-2 forms), college attendance (1098-T forms), and mortgage interest payment (1098 forms). The ZIP code on the claiming parent's 1040 return is typically from 1996 and is thus contingent on the ex post outcome of the student not having moved far from her elementary school for most students in our analysis sample.

Chetty et al. (2011) show that the match algorithm outlined below yields accurate matches for approximately 99% of cases in a school district sample that can be matched on social security number. Note that identifiers were used solely for the matching procedure. After the match was completed, the data were de-identified (i.e., individual identifiers such as names were stripped) and the statistical analysis was conducted using the de-identified dataset.

Step 1 [Date of Birth, Gender, Last Name]: We begin by matching each individual from the school-district data to Social Security Administration (SSA) records. We match individuals based on exact date of birth, gender, and the first four characters of last name. We only attempt to match individuals for which the school records include a valid date of birth, gender, and at least one valid last name. SSA records all last names ever associated in their records with a given individual; in addition, there are as many as three last names for each individual from the school files. We keep a potential match if any of these three last names match any of the last names present in the SSA file.

Step 2 [Rule Out on First Name]: We next check the first name (or names) of individuals from the school records against information from W2 and other information forms present in the tax records. Since these files reflect economic activity usually after the completion of school, we use this information in Step 2 only to "rule out" possible matches in order to minimize selection bias. In particular, we disqualify potential matches if none of the first names on the information returns match any of the first names in the school data. As before, we use only the first four characters of a first name. For many potential matches, we find no first name information in the tax information records; at this step we retain these potential matches. After removing potential matches that are mismatched on first name, we isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality (MQ) of matches identified at this stage as $MQ = 1$.

Step 3 [Dependent ZIP code]: For each potential match that remains, we find the household that claimed the individual as a dependent (if the individual was claimed at all) in each year. We then match the location of the claiming household, identified by the 5-digit ZIP code, to the home address ZIP code recorded in the school files. We classify potential matches based on the best ZIP code match across all years using the following tiers: exact match, match within 10 (e.g., 02139 and 02146 would qualify as a match), match within 100, and non-match. We retain potential matches only in the highest available tier of ZIP code match quality. For example, suppose there are 5 potential matches for a given individual, and that there are no exact matches on ZIP code, two matches within 10, two matches within 100, and one non-match. We would retain only the two that matched within 10. After this procedure, we isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match

pool. We classify the match quality of matches identified at this stage as $MQ = 2$.

Step 4 [Place of Birth]: For each potential match that remains, we match the state of birth from the school records with the state of birth as identified in SSA records. We classify potential matches into three groups: state of birth matches, state of birth does not match but the SSA state is the state where the school district is, and mismatches. Note that we include the second category primarily to account for the immigrants in the school data for whom the recorded place of birth is outside the country. For such children, the SSA state-of-birth corresponds to the state in which they received the social security number, which is often the first state in which they lived after coming to the country. We retain potential matches only in the best available tier of place-of-birth match quality. We then isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality of matches identified at this stage as $MQ = 3$.

Step 5 [Rule In on First Name]: After exhausting other available information, we return to the first name. To recall, in step 2 we retained potential matches that either matched on first name or for which there was no first name available. In this step, we retain only potential matches that match on first name, if such a potential match exists for a given student. We also use information on first name present on 1040 forms filed by potential matches as adults to identify matches at this stage. We then isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality of matches identified at this stage as $MQ = 4$.

Step 6 [Fuzzy Date-of Birth]: In previous work (Chetty et al. 2011), we found that 2-3% of individuals had a reported date of birth that was incorrect. In some cases the date was incorrect only by a few days; in others the month or year was off by one, or the transcriber transposed the month and day. To account for this possibility, we take all individuals for whom no eligible matches remained after step 2. Note that if any potential matches remained after step 2, then we would either settle on a unique best match in the steps that follow or find multiple potential matches even after step 5. We then repeat step 1, matching on gender, first four letters of last name, and fuzzy date-of-birth. We define a fuzzy DOB match as one where the absolute value of the difference between the DOB reported in the SSA and school data was in the set $\{1, 2, 3, 4, 59, 10, 18, 27\}$ in days, the set $\{1, 2\}$ in months, or the set $\{1\}$ in years. We then repeat steps 2 through 5 exactly as above to find additional matches. We classify matches found using this fuzzy-DOB algorithm as $MQ = 5.X$, where X is the corresponding MQ from the non-fuzzy DOB algorithm. For instance, if we find a unique fuzzy-DOB match in step 3 using dependent ZIP codes, then $MQ = 5.2$.

The following table shows the distribution of match qualities for all student-test-score observations. In all, we match 89.2% of student-subject observations in the analysis sample. We match 90.0% of observations in classes for which we are able to estimate VA for the teacher. Unmatched students are split roughly evenly among those for whom we found multiple matches and those for whom we found no match.

Match Quality (MQ)	Frequency	Percent	Cumulative Match Rate
1	3327727	55.63%	55.63%
2	1706138	28.52%	84.15%
3	146256	2.44%	86.59%
4	64615	1.08%	87.67%
5.1	84086	1.41%	89.08%
5.2	6450	0.11%	89.19%
5.3	747	0.01%	89.20%
5.4	248	0.00%	89.20%
Multiple Matches	304436	5.09%	
No Matches	341433	5.71%	

Appendix B: Identifying Teachers' Net Impacts

This appendix shows that the iterative method described in Section 6.1 recovers the net impacts of teacher VA, $\tilde{\beta}_g$, defined as the impact of raising teacher VA in grade g on earnings, holding fixed VA in subsequent grades.

We begin by estimating the following equations using OLS for $g \in [4, 8]$:

$$(12) \qquad Y_i = \beta_g \hat{\mu}_{j(i,g)} + f^{\mu}_{1g}(A_{i,t-1}) + f^{\mu}_{2g}(e_{j(i,g,t)}) + \phi^{\mu}_{1g} X_{igt} + \phi^{\mu}_{2g} \bar{X}_{c(i,g,t)} + \varepsilon^{\mu}_{igt}$$

$$(13) \; \hat{\mu}_{j(i,g')} = \rho_{gg'} \hat{\mu}_{j(i,g)} + f^{g'}_{1g}(A_{i,t-1}) + f^{g'}_{2}(e_{j(i,g,t)}) + \phi^{g'}_{1g} X_{igt} + \phi^{g'}_{2g} \bar{X}_{c(i,g,t)} + \eta_{itgg'} \quad \forall g' > g$$

The first set of equations estimates the reduced form impact of teacher VA in grade g on earnings. The second set of equations estimates the impact of teacher VA in grade g on teacher VA in future grade g'. Denote by $\mathbb{X}$ the vector of controls in equations (12) and (13). Note that identification of the tracking coefficients $\rho_{gg'}$ using (6.1) requires the following variant of Assumption 2:

Assumption 2A Teacher value-added in grade g is orthogonal to unobserved determinants of future teacher value-added:

$$Cov\left(\hat{\mu}_{j(i,g)}, \eta_{itgg'} \mid \mathbb{X}\right) = 0.$$

After estimating $\{\beta_g\}$ and $\{\rho_{gg'}\}$, we recover the net impacts $\tilde{\beta}_g$ as follows. Under our definition of $\tilde{\beta}_g$, earnings can be written as $\sum_{g=1}^{G} \tilde{\beta}_g \hat{\mu}_{j(i,g)} + \varepsilon^{\mu}_i$. Substituting this definition of Y_i into (12) and noting that $\rho_{gg'} = Cov\left(\hat{\mu}_{j(i,g')}, \hat{\mu}_{j(i,g)} \mid \mathbb{X}\right) / Var\left(\hat{\mu}_{j(i,g)} \mid \mathbb{X}\right)$ yields

$$\beta_g = \frac{Cov\left(\sum_{g'=1}^{G} \tilde{\beta}_{g'} \hat{\mu}_{j(i,g')} + \varepsilon^{Y}_i, \hat{\mu}_{j(i,g)} \mid \mathbb{X}\right)}{Var\left(\hat{\mu}_{j(i,g)} \mid \mathbb{X}\right)} = \sum_{g'=1}^{G} \rho_{gg'} \tilde{\beta}_{g'}.$$

One implication of Assumption 2, the orthogonality condition needed to identify earnings impacts, is that

$$Cov\left(\hat{\mu}_{j(i,g')}, \hat{\mu}_{j(i,g)} \mid \mathbb{X}\right) = 0 \quad \text{for } g' < g$$

since past teacher quality $\hat{\mu}_{j(i,g')}$ is one component of the error term ε^{μ}_{igt} in (12). Combined with

the fact that $\rho_{gg} = 1$ by definition, these equations imply that

$$\begin{aligned}
\beta_g &= \tilde{\beta}_g + \sum_{g'=g+1}^{G} \rho_{gg'} \tilde{\beta}_{g'} \quad \forall g < G \\
\beta_G &= \tilde{\beta}_G.
\end{aligned}$$

Rearranging this triangular set of equations yields the following system of equations, which can be solved by iterating backwards as in Section 6.1:

$$\begin{aligned}
\tilde{\beta}_G &= \beta_G \\
\tilde{\beta}_g &= \beta_g - \sum_{g'=g+1}^{G} \rho_{gg'} \tilde{\beta}_{g'} \quad \forall g < G.
\end{aligned} \tag{14}$$

FIGURE 1
Effects of Teacher Value-Added on Actual, Predicted, and Lagged Scores

Notes: These figures plot student scores, scaled in standard deviation units, vs. our leave-year-out measure of teacher value-added, which is also scaled in units of student test score standard deviations. The figures are drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by the same teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. In Panel A, the y variable is actual end-of-grade student scores; in Panel B, it is the predicted score based on parent characteristics; and in Panel C, it is the score two years before in the same subject. Predicted score is based on the fitted values from a regression of test score on mother's age at child's birth, indicators for parent's 401(k) contributions and home ownership, and an indicator for the parent's marital status interacted with a quartic in parent's household income (see Section 4.3 for details). All three figures control for the following classroom-level variables: school year and grade dummies, class-type indicators (honors, remedial), class size, and cubics in class and school-grade means of lagged test scores in math and English each interacted with grade. They also control for class and school-year means of the following student characteristics: ethnicity, gender, age, lagged suspensions, lagged absences, and indicators for grade repetition, special education, limited English. We use this baseline control vector in all subsequent figures unless otherwise noted. To construct each binned scatter plot, we first regress both the y- and x-axis variable on the control vector and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the x-axis residual and scatter the means of the y- and x-axis residuals within each bin. The solid line shows the best linear fit estimated on the underlying micro data estimated using OLS. The coefficients show the estimated slope of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.

FIGURE 2
Impacts of Teacher Value-Added on Lagged, Current, and Future Test Scores

Notes: This figure shows the effect of teacher value-added in year $t = 0$ on student scores from four years prior to assignment to the teacher of interest to four years after. The figure is drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Each point shows the coefficient estimate from a separate OLS regression of test scores (including all available grades and subjects) on teacher value-added and the baseline control vector used in Figure 1. The points for $t < -1$ represent placebo tests for selection on observables, while points for $t > 0$ show the persistence of teachers' impacts on test scores. The point at $t = 0$ corresponds to the regression coefficient in Panel A of Figure 1. The point at $t = -1$ is equal to zero by construction, because we control for lagged test scores. Teacher value-added is estimated using data from classes taught by the same teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. The coefficients from the regressions along with their associated standard errors are reported in Appendix Table 9.

FIGURE 3
Impacts of Teacher Entry and Exit on Average Test Scores by Cohort

Notes: These figures plot event studies of current scores (solid line) and prior-year scores (dashed line) by cohort as teachers enter or leave a school-grade-subject cell in year $t = 0$. Panels A and B analyze the entry and exit of a high-VA teacher (teachers with VA in the top 5% of the distribution); Panels C and D analyze the entry and exit of a low-VA (bottom 5%) teacher. All panels are plotted using a dataset containing school x grade x subject x year means from the linked analysis sample described in section 3.3. To construct each panel, we first estimate each teacher's VA using data from classes taught outside the years $t \in [-3, 2]$. We then plot mean scores in the subject taught by the teacher for students in the entire school-grade-subject cell in the years before and after the arrival or departure of the teacher. We remove year fixed effects by regressing the y variable on year indicators and plotting the mean of the residuals, adding back the sample mean of each variable to facilitate interpretation of the scale. Each point therefore shows the mean score of a different cohort of students within a single school-grade-subject cell, removing secular time trends. Each panel reports the change in mean score gains (mean scores minus mean lag scores) from $t = -1$ to $t = 0$. We also report the change in mean teacher VA multiplied by 0.861, the cross-class coefficient of score on VA (Column 1 of Table 2). We multiply the change in mean VA by this factor to forecast the change in test scores implied by the change in mean VA. We report p values from F tests of the hypotheses that the change in score gains from $t = -1$ to $t = 0$ equals 0 and equals the change in mean VA times 0.861. Mean teacher VA is calculated using a student-weighted average, imputing the sample mean for teachers who do not have data outside the $t \in [-3, 2]$ window.

FIGURE 4
Effect of Changes in Teaching Staff on Scores Across Cohorts

Notes: This figure plots changes in average test scores across cohorts versus changes in average teacher VA across cohorts, generalizing the event study in Figure 3 to include all changes in teaching staff. All panels are plotted using a dataset containing school x grade x subject x year means from the linked analysis sample described in section 3.3. We calculate changes in mean teacher VA across consecutive cohorts within a school-grade-subject cell as follows. First, we calculate teacher value-added for each teacher in a school-grade-subject cell in each adjacent pair of school years using information excluding those two years. We then calculate mean value-added across all teachers, weighting by the number of students they teach and imputing the sample mean VA for those for teachers for whom we have no estimate of VA. Finally, we compute the difference in mean teacher VA (year t minus year $t-1$) to obtain the x axis variable. The y axis variables are defined by calculating the change in the mean of the dependent variable (year t minus year t-1) within a school-grade-subject cell. In Panel A, the y-axis variable is the change in end-of-grade scores across cohorts in the relevant subject. In Panel B, the y-axis variable is the change in predicted test scores based on parent characteristics, defined as Figure 1b. In Panels C and D, the y-axis variable is the change in test scores in the other subject (e.g. math scores when analyzing English teachers' VA) for observations in elementary and middle school, respectively. To construct each binned scatter plot, we first regress both the y- and x-axis variable on year dummies and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the x-axis residual and scatter the means of the y- and x-axis residuals within each bin. The solid line shows the best linear fit estimated on the underlying school-grade-subject-year data estimated using an unweighted OLS regression. The coefficients show the estimated slope of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.

FIGURE 5
Effects of Teacher Value-Added on College Attendance

Notes: Panel A plots the relationship between teacher VA and college attendance rates at age 20. College attendance is measured by receipt of a 1098-T form in the year during which a student turned 20. The figure is drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. To construct the binned scatter plot, we first regress both the x- and y-variables on the baseline control vector used in Figure 1 and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the residual of the x variable and scatter the means of the y- and x-variable residuals within each bin, adding back the sample means of both variables to facilitate interpretation of the scale. The solid line shows the best linear fit estimated on the underlying micro data estimated using OLS. The coefficient shows the estimated slope of the best-fit line, with the standard error clustered at the school-cohort level reported in parentheses. Panel B replicates Panel A, changing the y variable to our earnings-based index of college quality at age 20. College quality is constructed using the average wage earnings at age 30 in 2009 for all students attending a given college at age 20 in 1999. For individuals who did not attend college, we calculate mean wage earnings at age 30 in 2009 for all individuals in the U.S. aged 20 in 1999 who did not attend any college. Panel C replicates the regression specification in Panel A and plots the resulting coefficients on college attendance from ages from 18 to 27. Each point represents the coefficient estimate on teacher value-added from a separate regression. The dashed lines show the boundaries of the 95% confidence intervals for the effect of value-added on college attendance at each age.

FIGURE 6
Effect of Teacher Value-Added on Earnings at Age 28

Notes: This figure plots the effect of teacher value-added on wage earnings at age 28, computed using data from W-2 forms issued by employers. The figure is drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. To construct the binned scatter plot, we first regress both earnings and value-added on the baseline control vector used in Figure 1 and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the value-added residual and scatter the means of the earnings and value-added residuals within each bin, adding back the sample means of earnings and value-added to facilitate interpretation of the scale. The solid line shows the best linear fit estimated on the underlying micro data estimated using OLS. The coefficient shows the estimated slope of the best-fit line, with the standard error clustered at the school-cohort level reported in parentheses.

FIGURE 7
Effect of Teacher Value-Added on Earnings by Age

Notes: This figure shows the effect of a 1 SD increase in teacher value-added on earnings at each age, expressed as a percentage of mean earnings at that age. The figure is drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. To construct the figure, we first run a separate OLS regression of earnings at each age (using all observations for which the necessary data are available) on teacher value-added, following exactly the specification used in Figure 7 . We then divide this regression coefficient by 10 to obtain an estimate of the impact of a 1 SD increase in teacher VA on earnings. Finally, we divide the rescaled coefficient by the mean earnings level in the estimation sample at each age to obtain the percentage impact of a 1 SD increase in VA on earnings by age. Panel A shows the results for the full sample. The dashed lines represent the 95% confidence interval, computed using standard errors clustered at the school-cohort level. Panel B replicates Panel A, splitting the sample into two based on the average college attendance rate at each school. The mean school-average college attendance rate is 35%. The solid series includes schools with attendance rates below 35% while the dashed series includes schools with attendance rates above 35%. The coefficients and standard errors underlying these figures are reported in Appendix Table 10.

FIGURE 8
Effects of Teacher Value-Added on Other Outcomes in Adulthood

Notes: These figures plot the effect of teacher value-added on teenage births (for females only) and neighborhood quality. We define a teenage birth as an individual claiming a dependent who was born when she was between the ages of 13 and 19 on the 1040 tax form in any year in our sample (see Section 3.2 for details). We define neighborhood quality as the fraction of residents with a college degree in the ZIP code where the individual resides. The figures are drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. To construct each binned scatter plot, we first regress both the y- and x-axis variables on the baseline control vector used in Figure 1 and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the x-axis residual and scatter the means of the y- and x-axis residuals within each bin, adding back the sample means of x- and y- axis variables to facilitate interpretation of the scales. The solid line shows the best linear fit estimated on the underlying micro data estimated using OLS. The coefficients show the estimated slopes of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.

FIGURE 9
Impacts of Teacher Value-Added on College Quality by Grade

Notes: This figure plots the impact of a 1 SD increase in teacher VA in each grade from 4-8 on our earnings-based index of college quality (projected earnings at age 30 based on the college in which the student is enrolled at age 20). The figure is drawn using the linked analysis sample described in section 3.3. The upper (circle) series shows the reduced-form effect of improved teacher quality in each grade, including both the direct impact of the teacher on earnings and the indirect effect through improved teacher quality in future years. Each point in this series represents the coefficient on teacher value-added from a separate regression of college quality at age 20 on teacher VA for a single grade. We use the same specification as in Figure 5c but limit the sample to cohorts who would have been in 4th grade during or after 1994 to obtained a balanced sample across grades. The shaded area represents a 95% confidence interval, calculated based on standard errors clustered by school-cohort. The lower (square) series plots the impact of teachers in each grade on college quality netting out the impacts of increased future teacher quality. We net out the effects of future teachers using the tracking coefficients reported in Appendix Table 13 and solving the system of equations in Section 6.1.

FIGURE 10
Earnings Impacts of Deselecting Low Value-Added Teachers

Notes: This figure displays the present value of lifetime earnings gains for a single classroom of students from deselecting teachers whose estimated value-added is in the bottom 5% of the distribution. The horizontal line shows the gain that could be achieved by deselecting the bottom 5% of teachers based on their true VA μ_j, measured noiselessly using an infinite number of classes per teacher. The increasing series plots the feasible gains from deselection of the bottom 5% of teachers when their VA is estimated based on the number of classes shown on the *x* axis, accounting for finite-sample error in VA estimates. Appendix Table 14 lists the values that are plotted as well as undiscounted cumulative earnings gains, which are approximately 5.5 times larger in magnitude. To obtain the values in the figure, we first calculate the present value of average lifetime earnings per student using the cross-sectional life-cycle earnings profile for the U.S. population in 2007, discounting earnings back to age 12 using a 3% net discount rate (equivalent to a 5% discount rate with 2% wage growth). Column 1 of Table 6 implies that a 1 SD increase in VA raises earnings by 0.9% at age 28. We assume that this 0.9% earnings gain remains constant over the life cycle and calculate the impacts of a 1 SD improvement in teacher quality on mean lifetime earnings, averaging across English and math teachers. Finally, we multiply the mean lifetime earnings impact by 28.3, the mean class size in our analysis sample.

APPENDIX FIGURE 1
Jacob and Levitt (2003) Proxy for Test Manipulation vs. Value-Added Estimates

Notes: This figure plots the relationship between our leave-out-year measure of teacher value added and Jacob and Levitt's proxy for cheating. The figure is drawn using the linked analysis sample described in section 3.3. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. The y- axis variable is constructed as follows. Let $\Delta\bar{A}_{c,t} = \bar{A}_{c,t} - \bar{A}_{c,t-1}$ denote the change in mean test scores from year $t-1$ to year t for students in classroom c. Let $R_{c,t}$ denote the ordinal rank of classroom c in $\Delta\bar{A}_{c,t}$ among classrooms in its grade, subject, and school year and $r_{c,t}$ the ordinal rank as a fraction of the total number of classrooms in that grade, subject, and school year. Jacob and Levitt's (2003) measure for cheating in each classroom is $JL_c = (r_{c,t})^2 + (1 - r_{c,t+1})^2$. Higher values of this proxy indicate very large test score gains followed by very large test score losses, which Jacob and Levitt show is correlated with a higher chance of having suspicious patterns of answers indicative of cheating. Following Jacob and Levitt, we define a classroom as an outlier if its value of JL_c falls within the top 5% of classrooms in the data. To construct the binned scatter plot, we group classrooms into percentiles based on their teacher's estimated value-added, ranking math and English classrooms separately. We then compute the fraction of Jacob-Levitt outliers within each percentile bin and scatter these fractions vs. the percentiles of teacher VA. Each point thus represents the fraction of Jacob-Levitt outliers at each subject-specific percentile of teacher VA, where VA for each teacher is estimated using data from other years. The dashed vertical line depicts the (subject-specific) 98th percentile of the value-added distribution. We exclude classrooms with estimated VA above this threshold in our baseline specifications because they have much higher frequencies of Jacob-Levitt outliers. See Appendix Table 8 for results with trimming at other cutoffs.

TABLE 1
Summary Statistics for Linked Analysis Dataset

Variable	Mean	S.D.	Observations
	(1)	(2)	(3)
Student Data:			
Class size (not student-weighted)	28.3	5.8	211,371
Number of subject-school years per student	6.14	3.16	974,686
Teacher experience (years)	8.08	7.72	4,795,857
Test score (SD)	0.12	0.91	5,312,179
Female	50.3%	50.0%	5,336,267
Age (years)	11.7	1.6	5,976,747
Free lunch eligible (1999-2009)	76.0%	42.7%	2,660,384
Minority (Black or Hispanic)	71.8%	45.0%	5,970,909
English language learner	10.3%	30.4%	5,813,404
Special education	3.4%	18.1%	5,813,404
Repeating grade	2.7%	16.1%	5,680,954
Student match rate to adult outcomes	89.2%	31.0%	5,982,136
Student match rate to parent chars.	94.6%	22.5%	5,329,715
Adult Outcomes:			
Annual wage earnings at age 20	4,796	6,544	5,255,599
Annual wage earnings at age 25	15,797	18,478	2,282,219
Annual wage earnings at age 28	20,327	23,782	851,451
In college at age 20	36.2%	48.1%	4,605,492
In college at age 25	17.3%	37.8%	1,764,179
College Quality at age 20	24,424	12,834	4,605,492
Contribute to a 401(k) at age 25	14.8%	35.5%	2,282,219
ZIP code % college graduates at age 25	13.2%	7.1%	1,919,115
Had a child while a teenager (for women)	8.4%	27.8%	2,682,644
Parent Characteristics:			
Household income (child age 19-21)	35,476	31,080	4,396,239
Ever owned a house (child age 19-21)	32.5%	46.8%	4,396,239
Contributed to a 401k (child age 19-21)	25.1%	43.3%	4,396,239
Ever married (child age 19-21)	42.1%	49.4%	4,396,239
Age at child birth	27.6	7.4	4,917,740
Predicted Score	0.16	0.26	4,669,069

Notes: All statistics reported are for the linked analysis dataset described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. The sample has one observation per student-subject-school year. Student data are from the administrative records of a large urban school district in the U.S. Adult outcomes and parent characteristics are from 1996-2010 federal income tax data. All monetary values are expressed in real 2010 dollars. All ages refer to the age of an individual as of December 31 within a given year. Teacher experience is the number of years of experience teaching in the school district. Test score is based on standardized scale scores, as described in Section 3.1. Free lunch is an indicator for receiving free of reduced-price lunches. Earnings are total wage earnings reported on W-2 forms, available from 1999-2010; those who are matched to tax data but have no W-2 are coded as having zero earnings. College attendance is measured by the receipt of a 1098-T form, available from 1999-2009. For a given college, "college quality" is defined as the average wage earnings at age 30 in 2009 for the subset of the U.S. population enrolled in that college at age 20 in 1999. For individuals who do not attend college, college quality is defined as the mean earnings at age 30 in 2009 of all individuals in the U.S. population not in college at age 20 in 1999. 401(k) contributions are reported on W-2 forms. ZIP code of residence is taken from either the address reported on 1040 or W-2 forms; for individuals without either in a given year, we impute location forward from the most recent non-missing observation. Percent college graduates in the ZIP code is based on data from the 2000 Census. Teenage births are measured only for females, by the claiming of a dependent, at any time in our sample, who was born when the claiming parent was between 13 and 19 years old. We link students to their parents by finding the earliest 1040 form from 1998-2010 on which the student is claimed as a dependent. We are unable to link 5.4% of matched students to their parents; the summary statistics for parents exclude these observations. Parent income is average adjusted gross income during the three tax-years when a student is aged 19-21. For parents who do not file, household income is defined as zero. Home ownership is measured by reporting mortgage interest payments on a 1040 or 1099 form. Marital status is measured by whether the claiming parent files a joint return while the child is between 19 and 21. Predicted score is predicted from a regression of scores on parent characteristics using the estimating equation in Section 4.3.

TABLE 2
Tests for Balance Using Parent Characteristics and Lagged Scores

Dep. Var.:	Score in year t	Predicted Score	Score in year t	Score in year t	Score in year t-2	Score in year t	Score in year t	Percent Matched
	(SD)	(SD)	(SD)	(SD)	(SD)	(SD)	(SD)	(%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher VA	0.861	0.006	0.866	0.864	-0.002	0.803	0.804	0.160
	(0.010)	(0.004)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.280)
	[82.68]	[1.49]	[75.62]	[75.85]	[-0.21]	[72.74]	[70.63]	[0.562]
Pred. score using par. chars.				0.175				
				(0.012)				
				[62.70]				
Year t-2 Score							0.521	
							(0.001)	
							[363.3]	
Observations	3,721,120	2,877,502	2,877,502	2,877,502	2,771,865	2,771,865	2,771,865	4,018,504

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses and t-statistics in square brackets. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher VA is scaled in units of student test score standard deviations. VA is estimated using data from classes taught by the same teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. In this and all subsequent tables, we exclude outlier observations with teacher VA in the top 2% of the VA distribution unless otherwise noted. In columns 1, 3, 4, 6, and 7, the dependent variable is the student's test score in a given year and subject. In column 2, the dependent variable is the predicted value generated from a regression of test score on mother's age at child's birth, indicators for parent's 401(k) contributions and home ownership, and an indicator for the parent's marital status interacted with a quartic in parent's household income. See Section 4.3 for details of the estimating equation for predicted scores. In column 5, the dependent variable is the score two years earlier in the same subject. The dependent variable in column 8 is an indicator for being matched to the tax data. The second independent variable in each of columns 4 and 7 is the same as the dependent variables in columns 2 and 5, respectively. All specifications control for the following classroom-level variables: school year and grade dummies, class-type indicators (honors, remedial), class size, and cubics in class and school-grade means of lagged test scores in math and English each interacted with grade. They also control for class and school-year means of the following student characteristics: ethnicity, gender, age, lagged suspensions, lagged absences, and indicators for grade repetition, special education, limited English. We use this baseline control vector in all subsequent tables unless otherwise noted.

TABLE 3
Sensitivity of Teacher Value-Added Measures to Controls

	(1) baseline	(2) add parent chars.	(3) add t-2 scores	(4) t-1 scores only	(5) no controls	(6) Quasi-Experimental Estimate of Bias
(1) baseline	1.000					3.1% (7.6)
(2) add parent chars.	0.999	1.000				2.6% (7.6)
(3) add t-2 scores	0.975	0.974	1.000			1.4% (7.4)
(4) t-1 scores only	0.945	0.943	0.921	1.000		14.3% (6.9)
(5) no controls	0.296	0.292	0.279	0.323	1.000	87.8% (1.4)

Notes: Columns 1-5 of this table report correlations between teacher value-added estimates from five models, each using a different control vector. The correlations are weighted by the number of years taught by each teacher. The models are estimated on a constant subsample of 89,673 classrooms from the linked analysis dataset for which the variables needed to estimate all five models are available. For each model, we estimate student test score residuals using equation (3) using the relevant control vector and then implement the remaining steps of the Empirical Bayes procedure in Section 2.2 identically. Model 1 uses the student- and class-level control vector used to estimate value-added in our baseline specifications. This control vector includes a cubic polynomial in prior-year scores in math and a cubic in prior-year scores in English interacted with the student's grade level, dummies for teacher experience, as well as the following student-level controls: ethnicity, gender, age, lagged suspensions and absences, and indicators for grade repetition, special education, limited English. The control vector also includes the following classroom-level controls: class-type indicators (honors, remedial), class size, cubics in class and school-grade means of lagged test scores in math and English each interacted with grade, class and school-year means of all the student-level controls, and school year and grade dummies. Model 2 adds classroom-level means of the following parental characteristics to model 1: parent's age at child's birth, mean parent household income, and indicators for whether the parent owned a house, invested in a 401k, or was married while child was 19-21, and an indicator for whether no parent was found for the child in the tax data. Model 3 adds a cubic in twice-lagged test scores in the same subject to model 1. Model 4 controls for only lagged scores, using cubics in student's prior-year math and English scores interacted with grade level and cubics in the mean prior-year math and English scores for the classroom and school-grade cell also interacted with grade level. Model 5 includes no controls. In column 6, we estimate the degree of bias in the VA estimates produced by each model using quasi-experimental changes in teaching staff as described in Section 4.4. To calculate the degree of bias, we first estimate the effect of changes in mean VA on changes in test scores across cohorts using the specification in Column 1 of Table 4. We then estimate the effect of differences in teacher VA across classrooms on test scores, using the specification in Column 1 of Table 2 but with the control vector corresponding exactly to that used to estimate the value-added model. Finally, we define the degree of bias as the percentage difference between the cross-cohort and cross-class coefficients. Standard errors for the bias calculation are calculated as the standard error of the coefficient in the cross-cohort regression divided by the cross-class estimate; this calculation ignores the error in the cross-class estimate, which is negligible, as shown in Column 1 of Table 2.

TABLE 4
Impacts of Quasi-Experimental Changes in Teaching Staff on Test Scores

Dependent Variable:	Δ Score	Δ Predicted Score	Δ Other Subj. Score	Δ Other Subj. Score
	(SD)	(SD)	(SD)	(SD)
	(1)	(2)	(3)	(4)
Changes in mean teacher VA across cohorts	0.843	0.008	0.694	0.005
	(0.053)	(0.011)	(0.058)	(0.105)
	[15.95]	[0.74]	[11.90]	[0.04]
Grades	4 to 8	4 to 8	Elem. Sch.	Middle Sch.
Number of school x grade x subject x year cells	24,887	25,073	20,052	4,651

Notes: Each column reports coefficients from an unweighted OLS regression, with standard errors clustered by school-cohort in parentheses and t-statistics in square brackets. The regressions are estimated on a dataset containing school x grade x subject x year means from the linked analysis sample described in section 3.3. We calculate changes in mean teacher VA across consecutive cohorts within a school-grade-subject cell as follows. First, we calculate teacher value-added for each teacher in a school-grade-subject cell in each adjacent pair of school years using information excluding those two years. We then calculate mean value-added across all teachers, weighting by the number of students they teach and imputing the sample mean VA for those for teachers for whom we have no estimate of VA. Finally, we compute the difference in mean teacher VA (year t minus year t-1) to obtain the independent variable. We do not exclude teachers whose estimated VA is in the top 2% of the distribution when computing mean VA. The dependent variables are defined by calculating the change in the mean of the dependent variable (year t minus year t-1) within a school-grade-subject cell. In column 1, the dependent variable is the change in mean scores in the corresponding subject. In column 2, it is the change in the predicted score, constructed as described in the notes to Table 2. In Columns 3 and 4, the dependent variable is the change in the score in the other subject (e.g. math scores for English teachers). Column 3 restricts the sample to elementary schools, where math and English are taught by the same teacher; column 4 restricts the sample to middle schools, where different teachers teach the two subjects. All specifications include no controls except year fixed effects.

TABLE 5
Impacts of Teacher Value-Added on College Attendance

Dep. Var.:	College at Age 20	Pred. College at Age 20	Changes in Age 20 Coll. Attendance	College Quality at Age 20	Changes in Age 20 Coll. Quality	High Quality Coll. at Age 20	College at Age 25
	(%)	(%)	(%)	($)	($)	(%)	(%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Teacher VA	4.917	0.463		1,644		3.588	2.752
	(0.646)	(0.261)		(173)		(0.612)	(0.697)
Changes in mean VA across cohorts			6.101		1,319		
			(2.094)		(539)		
Controls	x	x		x		x	x
Source of Variation	X-Class	X-Class	X-Cohort	X-Class	X-Cohort	X-Class	X-Class
Observations	3,095,822	3,097,322	25,073	3,095,822	24,296	3,095,822	985,500
Mean of Dep. Var.	37.8	37.8	35.9	24,815	24,293	19.8	18.1

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. Columns 1, 2, 4, 6, and 7 use cross-class variation, while columns 3 and 5 use cross-cohort variation. For specifications that use cross-class variation, teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the baseline control vector in model 1 of Table 3. The dependent variable in column 1 is an indicator for college attendance at age 20. The dependent variable in column 2 is the predicted value generated from a regression of college attendance at age 20 on parent characteristics, using the same specification as for predicted score described in the notes to Table 2. The dependent variable in column 4 is the earnings-based index of college quality, defined in the notes to Table 1. The dependent variable in column 6 is an indicator for attending a college whose quality is greater than the median college quality among those attending college, which is $39,972. The dependent variable in column 7 is an indicator for college attendance at age 25. All cross-class regressions include the baseline class-level control vector used in Table 2. For the specifications that exploit cross-cohort variation in columns 3 and 5, we use changes in mean teacher value-added as the main independent variable, defined exactly as in Table 4. The dependent variables in Columns 3 and 5 are changes in mean college attendance and quality across consecutive cohorts within a school-grade-subject cell. Columns 3 and 5 include no controls except year fixed effects.

TABLE 6
Impacts of Teacher Value-Added on Earnings

Dep. Var.:	Earnings at Age 28	Earnings at Age 30	Wage Growth Ages 22-28	College at Age 25	College at Age 25	Wage Growth Ages 22-28	Wage Growth Ages 22-28
	($)	($)	($)	(%)	(%)	($)	($)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Teacher VA	1,815	2,058	1,802	0.526	4.728	1,403	2,838
	(729)	(1953)	(636)	(0.789)	(1.152)	(661)	(1,118)
Observations	368,427	61,639	368,405	528,065	457,435	201,933	166,472
Schools	All	All	All	Low Coll.	High Coll.	Low Coll.	High Coll.
Mean of Dep. Var.	20,912	22,347	14,039	14.30	22.43	10,159	18,744

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. The dependent variable in columns 1 and 2 are the individual's wage earnings reported on W-2 forms at ages 28 and 30, respectively. The dependent variable in columns 3, 6, and 7 is the change in wage earnings between ages 22 and 28. The dependent variable in columns 4 and 5 is an indicator for attending college at age 25. All regressions exploit variation across classrooms and include the baseline class-level control vector used in Table 2. Columns 1-3 use the entire analysis sample. In columns 4-7, we split the sample into two based on the average college attendance rate at each school. The mean school-average college attendance rate is 35%. Columns 4 and 6 use schools with attendance rates below 35% while columns 5 and 7 use schools with attendance rates above 35%.

TABLE 7
Impacts of Teacher Value-Added on Other Outcomes

Dep. Var.:	Teenage Birth	Percent College Grads in ZIP at Age 25	Percent College Grads in ZIP at Age 28	401(k) at Age 25	401(k) at Age 25
	(%)	(%)	(%)	(%)	(%)
	(1)	(2)	(3)	(4)	(5)
Value-Added	-0.991	0.628	1.439	1.885	-1.780
	(0.353)	(0.194)	(0.310)	(0.680)	(0.987)
Observations	1,826,742	1,168,965	310,638	725,140	646,955
Schools	All	All	All	Low Coll.	High Coll.
Mean of Dep. Var.	7.9	13.3	13.6	12.1	19.2

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. The dependent variable in column 1 is an indicator for having a teenage birth, defined as in Table 1. The dependent variable in columns 2 and 3 is the fraction of residents in an individual's zip code of residence at ages 25 and 28 with a college degree or higher, based on data from the 2000 Census. ZIP code is obtained from either 1040 or W-2 forms filed in the current year or imputed from past years for non-filers. The dependent variable in columns 4 and 5 is an indicator for whether an individual made a contribution to a 401(k) plan at age 25. All regressions exploit variation across classrooms and include the baseline class-level control vector used in Table 2. Columns 1-3 use the entire analysis sample. In columns 4 and 5, we split the sample into two based on the average college attendance rate at each school. The mean school-average college attendance rate is 35%. Columns 4 and 6 use schools with attendance rates below 35% while columns 5 and 7 use schools with attendance rates above 35%.

TABLE 8
Heterogeneity in Impacts of Teacher Value-Added

Panel A: Impacts by Demographic Group

	Girls	Boys	Low Income	High Income	Minority	Non-Minority
	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable: College Quality at Age 20 ($)					
Teacher VA	1,903	1,386	1,227	2,087	1,302	2,421
	(211)	(203)	(174)	(245)	(154)	(375)
Mean of Dep. Var.	25,509	24,106	21,950	27,926	21,925	31,628
	Dependent Variable: Test Score (SD)					
Teacher VA	0.856	0.863	0.843	0.865	0.846	0.889
	(0.012)	(0.013)	(0.014)	(0.013)	(0.012)	(0.018)
Mean of Dep. Var.	0.191	0.161	-0.010	0.324	-0.037	0.663

Panel B: Impacts by Subject

	Elementary School			Middle School		
	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable: College Quality at Age 20 ($)					
Math Teacher VA	1095		638	1,648		1,374
	(176)		(219)	(357)		(347)
English Teacher VA		1,901	1,281		2,896	2,543
		(303)	(376)		(586)	(574)

Notes: Each cell reports a coefficient from a separate OLS regression of an outcome on teacher value-added, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. All regressions exploit variation across classrooms and include the baseline class-level control vector used in Table 2. In Panel A, there is one observation for each student-subject-school year; in Panel B, the data are reshaped so that both subjects (math and English) are in the same row, with one observation for each student-school year. The dependent variable in the top half of Panel A and in Panel B is the earnings-based index of college quality (see Table 1 for details). The dependent variable in the second half of Panel A is the student's test score. In Panel A, we split the sample in columns 1 and 2 between boys and girls. We split the sample in columns 3 and 4 based on whether a student's parental income is higher or lower than median in sample, which is $26,961. We split the sample in columns 5 and 6 based on whether a student belongs to an ethnic minority (Black or Hispanic). In Panel B, we split the sample into elementary schools (schools where the student is taught by the same teacher for both math and English) and middle schools (which have different teachers for each subject). All specifications in Panel B control for the baseline class-level variables described in Table 2 in both the student's math and English classrooms.

APPENDIX TABLE 1
Structure of Linked Analysis Dataset

Student	Subject	Year	Grade	Class	Teacher	Test Score	Matched to Tax Data?	In college at Age 20?	Earnings at Age 28	Parent Income
			…							
Bob	Math	1992	4	1	Jones	0.5	1	1	$27K	$95K
Bob	English	1992	4	1	Jones	-0.3	1	1	$27K	$95K
Bob	Math	1993	5	2	Smith	0.9	1	1	$27K	$95K
Bob	English	1993	5	2	Smith	0.1	1	1	$27K	$95K
Bob	Math	1994	6	3	Harris	1.5	1	1	$27K	$95K
Bob	English	1994	6	4	Adams	0.5	1	1	$27K	$95K
Nancy	Math	2002	3	5	Daniels	0.4	0	.	.	
Nancy	English	2002	3	5	Daniels	0.2	0	.	.	
Nancy	Math	2003	4	6	Jones	-0.1	0	.	.	
Nancy	English	2003	4	6	Jones	0.1	0	.	.	
			…							

Notes: This table illustrates the structure of the analysis dataset, which combines information from the school district database and the tax data. The linked analysis data includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one row for each student-subject-school year, with 5,928,136 rows in total. Individuals who were not linked to the tax data have missing data on adult outcomes and parent characteristics. The values in this table are not real data and for illustrative purposes only.

APPENDIX TABLE 2
Summary Statistics for School District Data Used to Estimate Value-Added

Variable	Mean (1)	S.D. (2)	Observations (3)
Class size (not student-weighted)	27.5	5.1	318,812
Number of subject-school years per student	5.58	2.98	1,375,552
Teacher Experience (years)	7.3	7.3	7,675,495
Test score (SD)	0.17	0.88	7,675,495
Female	50.8%	50.0	7,675,288
Age (years)	11.4	1.5	7,675,282
Free lunch eligible (1999-2009)	79.6%	40.3%	5,046,441
Minority (Black or Hispanic)	71.6%	45.1%	7,672,677
English language learner	5.1%	22%	7,675,495
Special education	1.9%	13.7%	7,675,495
Repeating grade	1.7%	13.0%	7,675,495

Notes: Statistics reported are for the set of observations used to estimate teacher value-added. These are observations from the full school district dataset spanning 1991-2009 described in section 3.1 that have information on test scores, teachers, and all the control variables (such as lagged test scores) needed to estimate the baseline value-added model in Table 3. We exclude observations from classrooms that have fewer than 7 students with the necessary information to estimate value-added. The sample has one observation per student-subject-school year. See notes to Table 1 for definitions of variable and additional details.

APPENDIX TABLE 3
Cross-Sectional Correlations Between Outcomes in Adulthood and Test Scores

Dep. Var.:	Earnings at Age 28	College at Age 20	College Quality at Age 20	Teenage Birth	Percent College Grads in ZIP at Age 25
	($)	(%)	($)	(%)	(%)
	(1)	(3)	(2)	(4)	(5)
No Controls	7,601	18.33	6,030	-3.84	1.85
	(28)	(0.02)	(6)	(0.02)	(0.01)
With Controls	2,539	5.66	2,009	-1.03	0.37
	(76)	(0.05)	(13)	(0.04)	(0.01)
Math	2,813	5.97	2,131	-0.88	0.34
Full Controls	(104)	(0.07)	(18)	(0.06)	(0.02)
English	2,194	5.27	1,843	-1.21	0.38
Full Controls	(112)	(0.07)	(18)	(0.06)	(0.02)
Mean of Dep. Var.	20,867	37.2	24,678	8.25	13.2

Notes: Each cell reports coefficients from a separate OLS regression of an outcome in adulthood on test scores measured in standard deviation units, with standard errors reported in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year, and we pool all subjects and grades in estimating these regressions. The dependent variable is wage earnings at age 28 in column 1, an indicator for attending college at age 20 in column 2, our earnings-based index of college quality in column 3, an indicator for having a teenage birth (defined for females only) in column 4, and the fraction of residents in an individual's zip code of residence with a college degree or higher at age 25 in column 5. See notes to Table 1 for definitions of these variables. The regressions in the first row include no controls. The regressions in the second row include the full vector of student- and class-level controls used to estimate the baseline value-added model, described in the notes to Table 3. The regressions in the third and fourth row both include the full vector of controls and split the sample into math and English test score observations.

APPENDIX TABLE 4
Cross-Sectional Correlations Between Test Scores and Earnings by Age

	Dependent Variable: Earnings ($)										
	20	21	22	23	24	25	26	27	28	29	30
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
No Controls	435	548	1,147	2,588	3,516	4,448	5,507	6,547	7,440	8,220	8,658
	(15)	(19)	(24)	(30)	(36)	(42)	(49)	(56)	(63)	(68)	(72)
With Controls	178	168	354	942	1,282	1,499	1,753	2,151	2,545	2,901	3,092
	(46)	(60)	(75)	(94)	(111)	(130)	(151)	(172)	(191)	(208)	(219)
Mean Earnings	4,093	5,443	6,986	9,216	11,413	13,811	16,456	19,316	21,961	23,477	23,856
Pct. Effect (with controls)	4.4%	3.1%	5.1%	10.2%	11.2%	10.9%	10.7%	11.1%	11.6%	12.4%	13.0%

Notes: Each cell in the first two rows reports coefficients from a separate OLS regression of earnings at a given age on test scores measured in standard deviation units, with standard errors in parentheses. We obtain data on earnings from W-2 forms and include individuals with no W-2's as observations with 0 earnings. The regressions are estimated on a constant subsample of the linked analysis sample, i.e. the subset of students for whom data on earnings are available from ages 20-30. There is one observation for each student-subject-school year, and we pool all subjects and grades in estimating these regressions. The first row includes no controls; the second includes the full vector of student- and class-level controls used to estimate the baseline value-added model, described in the notes to Table 3. Means of earnings for the available estimation sample are shown in the third row. The last row divides the coefficient estimates from the specification with controls by the mean earnings to obtain a percentage impact by age.

APPENDIX TABLE 5
Heterogeneity in Cross-Sectional Correlations Across Demographic Groups

Dependent Variable:	Earnings at Age 28	College at at Age 20	College Quality Age 20	Teenage Birth
	($)	(%)	($)	(%)
	(1)	(2)	(3)	(4)
Male	2,235	5.509	1,891	n/a
	(112)	(0.069)	(18)	n/a
	[21,775]	[0.34567]	[24,268]	n/a
Female	2,819	5.828	2,142	-1.028
	(102)	(0.073)	(19)	(0.040)
	[20,889]	[0.42067]	[25,655]	[0.07809]
Non-minority	2,496	5.560	2,911	-0.550
	(172)	(0.098)	(30)	(0.039)
	[31,344]	[0.60147]	[32,288]	[0.01948]
Minority	2,583	5.663	1,624	-1.246
	(80)	(0.058)	(13)	(0.053)
	[17,285]	[0.29627]	[22,031]	[0.10038]
Low Parent Inc.	2,592	5.209	1,571	-1.210
	(108)	(0.072)	(17)	(0.072)
	[17,606]	[0.27636]	[22,011]	[0.10384]
High Parent Inc.	2,614	5.951	2,414	-0.834
	(118)	(0.072)	(19)	(0.054)
	[26,688]	[0.49882]	[28,038]	[0.05974]

Notes: Each cell reports coefficients from a separate OLS regression of an outcome in adulthood on test scores measured in standard deviation units, with standard errors reported in parentheses. Means of the dependent variable for the relevant estimation sample are shown in square brackets. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year, and we pool all subjects and grades in estimating these regressions. The dependent variable is wage earnings at age 28 in column 1, an indicator for attending college at age 20 in column 2, our earnings-based index of college quality in column 3, and an indicator for having a teenage birth (defined for females only) in column 4. All regressions include the full vector of student- and class-level controls used to estimate the baseline value-added model, described in the notes to Table 3. The demographic groups are defined in exactly the same way as in Table 8. We split the sample in rows 3 and 4 based on whether a student belongs to an ethnic minority (Black or Hispanic). We split the sample in rows 5 and 6 based on whether a student's parental income is higher or lower than median in sample, which is $26,961.

APPENDIX TABLE 6

Cross-Sectional Correlations between Test Scores and Outcomes in Adulthood by Grade

Dep. Variable:	Earnings at Age 28	College at Age 20	College Quality at Age 20	Earnings at Age 28	College at Age 20	College Quality at Age 20
	No Controls			With Controls		
	($)	(%)	($)	($)	(%)	($)
	(1)	(2)	(3)	(4)	(5)	(6)
Grade 4	7,618	18.2	5,979	3,252	6.763	2,360
	(76.7)	(0.053)	(13.8)	(157)	(0.099)	(25.4)
Grade 5	7,640	18.3	6,065	2,498	5.468	1,994
	(61.6)	(0.052)	(13.6)	(129)	(0.096)	(24.8)
Grade 6	7,395	18.0	5,917	2,103	4.987	1,778
	(63.0)	(0.057)	(14.7)	(161)	(0.118)	(29.8)
Grade 7	7,790	18.4	5,950	2,308	4.844	1,667
	(64.6)	(0.060)	(15.5)	(342)	(0.133)	(33.2)
Grade 8	7,591	18.9	6,228	2,133	5.272	1,913
	(54.7)	(0.055)	(14.1)	(196)	(0.129)	(32.3)
Mean of Dep Var.	20,867	37.17	24,678	20,867	37.17	24,678

Notes: Each cell reports coefficients from a separate OLS regression of an outcome in adulthood on end-of-grade test scores measured in standard deviation units, using data from only a single grade. Standard errors are reported in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Columns 1-3 do not include any controls. Columns 4-6 include the full vector of student- and class-level controls used to estimate the baseline value-added model, described in the notes to Table 3. The dependent variable in columns 1 and 4 is wage earnings at age 28. The dependent variable in columns 2 and 5 is an indicator for college attendance at age 20. The dependent variable in columns 3 and 6 is our earnings-based index of college quality.

APPENDIX TABLE 7
Robustness Analysis: Clustering and Control Vectors

Dependent Variable:	Score	College at Age 20	Earnings at Age 28
	(SD)	(%)	($)
	(1)	(2)	(3)
Panel A: Baseline Analysis Sample			
Baseline estimates	0.861	0.049	1815
Baseline s.e. (school-cohort)	(0.010)	(0.006)	(727)
95% CI	(0.841, 0.882)	(0.037,0.062)	(391, 3240)
95% CI using student bootstrap	(0.851, 0.871)	(0.040, 0.056)	(630, 3095)
p value using student bootstrap	<.01	<.01	<.01
Panel B: Observations with Data on Earnings at Age 28			
	1.157	0.060	1815
no clustering	(0.016)	(0.010)	(531)
school-cohort	(0.036)	(0.016)	(727)
two-way student and class	(0.029)	(0.013)	(675)
Panel C: First observation for each child, by subject			
Math	0.986	0.040	1258
no clustering	(0.009)	(0.006)	(780)
school-cohort	(0.017)	(0.007)	(862)
class	(0.016)	(0.007)	(848)
English	1.116	0.061	2544
no clustering	(0.015)	(0.010)	(1320)
school-cohort	(0.025)	(0.012)	(1576)
class	(0.024)	(0.012)	(1516)
Panel D: Additional Controls			
Baseline class controls	0.858	0.049	1696
school-cohort	(0.010)	(0.007)	(797)
Add Individual Controls	0.856	0.049	1688
school-cohort	(0.010)	(0.007)	(792)
Add School-Year Effects	0.945	0.026	1942
school-cohort	(0.009)	(0.005)	(669)

Notes: This table reports coefficient estimates, with standard errors or confidence intervals in parentheses, from OLS regressions of various outcomes on teacher value-added. The dependent variable in column 1 is score. The dependent variable in column 2 is an indicator for college attendance at age 20. The dependent variable in column 3 is wage earnings at age 28. Panel A reports the results from the baseline specifications estimated on the full linked analysis sample, along with a 95% confidence interval generated from a block-bootstrap at the student level. Panel B reports results for the subsample of observations for whom we have data on earnings at age 28. We report three sets of standard errors: no clustering, clustering by school-cohort as in our baseline analysis, and two-way clustering by student and classroom (Cameron, Gelbach, and Miller 2011). In Panel C, we eliminate repeated observations at the individual level by using only the first observation per student in each subject. We then report the same three sets of standard errors. Finally, Panel D evaluates the sensitivity of the estimates to changes in the control vector. The first and second rows of Panel D use the subsample of observations with non-missing student-level controls. The first row uses the baseline classroom-level controls used in Table 2 and other tables, while the second adds the student-level controls used to estimate our baseline value-added model (model 1 in Table 3). The third row uses the full analysis sample and includes school-year fixed effects in both the estimation of teacher VA and the outcome regressions.

APPENDIX TABLE 8
Impacts of Teacher Value-Added: Sensitivity to Trimming

	Percent Trimmed in Upper Tail						Bottom and	Jacob and
	5%	4%	3%	**2%**	1%	0%	Top 2%	Levitt proxy
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test Score	0.846	0.853	0.860	0.861	0.868	0.870	0.866	0.754
	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
College at Age 20	5.724	5.585	5.258	4.917	4.730	4.022	4.091	6.455
	(0.693)	(0.673)	(0.662)	(0.646)	(0.622)	(0.590)	(0.668)	(0.703)
College Quality	1,870	1,848	1,773	1,644	1,560	1,432	1,425	2,068
at Age 20	(185)	(180)	(177)	(173)	(167)	(160)	(177)	(187)
Earnings at Age 28	2,058	2,080	1,831	1,815	1,581	994	1,719	1,672
	(808)	(776)	(745)	(729)	(709)	(668)	(797)	(834)

Notes: Each coefficient reports the coefficient on teacher VA from a separate OLS regression, with standard errors clustered by school-cohort in parentheses. The dependent variable is end-of-grade test score in the first row, an indicator for college attendance in the second row, our earnings-based index of college quality in the third row, and earnings at age 28 in the fourth row. The regressions in each of these rows replicate exactly the baseline cross-class specification used in Column 1 of Table 2, Columns 1 and 5 of Table 4, and Column 1 of Table 5. The baseline estimates are reported in column 4, which shows the results with trimming the top 2% of VA outliers. Columns 1-6 report results for trimming the upper tail at other cutoffs. Column 7 shows estimates when both the bottom and top 2% of VA outliers are excluded. Finally, column 8 excludes teachers who have more than one classroom that is an outlier according to Jacob and Levitt's (2003) proxy for cheating. Jacob and Levitt define an outlier classroom as one that ranks in the top 5% of a test-score change metric defined in the notes to Appendix Figure 1.

APPENDIX TABLE 9

Impacts of Teacher Value-Added on Lagged, Current, and Future Test Scores

	Dependent Variable: Test Score (SD)							
	t-4	t-3	t-2	t	t+1	t+2	t+3	t+4
	(1)	(2)	(3)	(5)	(6)	(7)	(8)	(9)
Teacher VA	-0.059	-0.041	-0.004	0.861	0.499	0.393	0.312	0.297
	(0.020)	(0.015)	(0.011)	(0.010)	(0.011)	(0.012)	(0.013)	(0.018)
Observations	1,184,397	1,906,149	2,826,978	3,721,120	2,911,042	2,247,141	1,578,551	790,173

Notes: This table reports the values plotted in Figure 2. Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. Each column replicates exactly the specification in Column 1 of Table 2, replacing the dependent variable with scores in year $t + s$ to measure the impact of teacher quality in year t, where s varies from -4 to 4. We omit the specification for $s = -1$ since we control for lagged score.

APPENDIX TABLE 10
Impacts of Teacher Value-Added on Earnings by Age

Dependent Variable:	Earnings ($)								
	20	21	22	23	24	25	26	27	28
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Panel A: Full Sample								
Value-Added	-211	-322	-211	71	449	558	1,021	1,370	1,815
	(72)	(100)	(136)	(190)	(247)	(319)	(416)	(517)	(729)
Mean Earnings	4,872	6,378	8,398	11,402	13,919	16,071	17,914	19,322	20,353
	Panel B: Schools with Low College Attendance Rates								
Value-Added	171	165	416	731	1,053	1,405	1,637	1,728	2,073
	(87)	(119)	(159)	(215)	(277)	(343)	(440)	(546)	(785)
Mean Earnings	4,747	6,183	7,785	9,752	11,486	13,064	14,319	15,249	15,967
	Panel C: Schools with High College Attendance Rates								
Value-Added	-592	-791	-870	-730	-318	-464	448	1,200	2,209
	(110)	(157)	(217)	(317)	(417)	(554)	(717)	(911)	(1,274)
Mean Earnings	5,018	6,609	9,127	13,379	16,869	19,774	22,488	24,718	26,312

Notes: Each coefficient reports the effect of teacher VA on earnings from a separate OLS regression, with standard errors clustered by school-cohort in parentheses. All regressions use the specification and sample used to estimate Column 1 of Table 5, replacing the dependent variable with earnings at the age shown in the column heading. In Panels B and C, we split the sample into two based on the average college attendance rate at each school. The mean school-average college attendance rate is 35%. Panel B considers schools with attendance rates below 35% while Panel C considers schools with attendance rates above 35%. The second row in each panel reports mean earnings for the observations in the corresponding estimation sample.

APPENDIX TABLE 11
Impacts of Teacher Quality: Instrumental Variables Specifications

Dependent Variable:	Score	College at Age 20	College Quality at Age 20	Earnings at Age 28	College at Age 20	College Quality at Age 20	Earnings at Age 28
Estimation Method:	OLS (Reduced Form)				Two-Stage Least Squares		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	(SD)	(%)	($)	($)	(%)	($)	($)
Raw Teacher Quality	0.476	2.526	837	871			
	(0.006)	(0.349)	(93)	(392)			
Score					5.29	1,753	1,513
					(0.72)	(191)	(673)
Observations	3,721,120	3,095,822	3,095,822	368,427	3,089,442	3,089,442	368,427
Mean of Dep. Variable	0.162	37.8	24,815	20,912	37.8	24,815	20,912

Notes: This table reproduces the baseline specifications in Table 2 (Col. 1), Table 4 (Cols. 1 and 4), and Table 5 (Col 1) using raw estimates of teacher quality. Raw teacher quality is the estimate ν_j obtained after Step 2 of the procedure described in Section 2.2, prior to the Empirical Bayes shrinkage correction. We define teacher quality using student score residuals from classes taught by the same teacher in all other years available in the school district dataset. Student score residuals are calculated from an OLS regression of scores on the full student- and classroom-level control vector used to estimate the baseline value-added model, defined in the notes to Table 3. Columns 1 through 4 regress the outcome on raw teacher quality with the baseline classroom-level control vector used in Table 2. Columns 5-7 report 2SLS estimates, instrumenting for mean classroom test scores with raw teacher quality. All regressions cluster standard errors at the school x cohort level and are estimated on the linked analysis sample used to estimate the baseline specifications, with one observation per student-subject-school year. For comparability to baseline estimates, observations with teacher VA in the top 2% of the distribution (estimated using the baseline model in Table 3) are excluded.

APPENDIX TABLE 12
Impacts of Value-Added on College Quality by Grade

	College Quality at Age 20				
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Panel A: Reduced-Form Coefficients					
Teacher Value-Added	2,011	832	788	2,638	970
	(296)	(314)	(363)	(472)	(398)
Panel B: Coefficients Net of Teacher Tracking					
Teacher Value-Added	1,991	802	566	2,478	970

Notes: This table reports the coefficients plotted in Figure 10. Panel A replicates Column 5 of Table 4 for each grade separately, using only cohorts who would have been in 4th grade during or after 1994. Panel B calculates the impacts of teacher VA in each grade net of tracking to better teachers in future grades. We obtain these point estimates by estimating the impact of VA on future VA (see Appendix Table 13) and then subtracting out the indirect effects using the procedure described in section 6.1.

APPENDIX TABLE 13
Tracking Coefficients

	Future Teacher Quality			
	Grade 5	Grade 6	Grade 7	Grade 8
Grade 4 Teacher VA	0.001	0.012	0.005	-0.001
Grade 5 Teacher VA		0.038	0.002	0.004
Grade 6 Teacher VA			0.067	0.058
Grade 7 Teacher VA				0.165

Notes: Each cell reports the coefficient from a separate regression of teacher value-added in a subsequent grade on teacher value-added in the current grade. All regressions include the classroom-level baseline control vector used in Table 2 and are estimated on the linked analysis sample, using all observations for which the data needed to estimate the relevant regression are available.

APPENDIX TABLE 14
Lifetime Earnings Impacts of Deselecting Teachers Below 5th Percentile

Num. of Classes Observed	Present Value at Age 12 of Earnings Gain per Class	Undiscounted Sum of Earnings Gain per Class
1	$135,228	$748,248
2	$169,865	$939,899
3	$189,247	$1,047,145
4	$201,917	$1,117,250
5	$210,923	$1,167,085
6	$217,683	$1,204,486
7	$222,955	$1,233,659
8	$227,188	$1,257,083
9	$230,665	$1,276,321
10	$233,574	$1,292,415
Max	$266,664	$1,475,511

Notes: This table shows the earnings gains from replacing a teacher whose value-added is in the bottom 5% of the distribution with a median teacher for a single class of average size (28.3 children). Column 1 reports present values of earnings gains, discounted back to age 12 at a 5% rate. Column 2 reports undiscounted sums of total earnings gains. The row labeled Max shows the gains from deselecting teachers based on their on true VA. The other rows show the gains from deselection when VA is estimated based on a given number of classes. The calculations are based on the average lifecycle income profile of individuals in the U.S. population in 2007, adjusted for a 2% annual growth rate in earnings.

EXHIBIT 2
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

MET project

POLICY AND PRACTICE BRIEF

Ensuring Fair and Reliable Measures of Effective Teaching

Culminating Findings from the MET Project's Three-Year Study

BILL & MELINDA GATES foundation

ABOUT THIS REPORT: This non-technical research brief for policymakers and practitioners summarizes recent analyses from the Measures of Effective Teaching (MET) project on identifying effective teaching while accounting for differences among teachers' students, on combining measures into composites, and on assuring reliable classroom observations.[1]

Readers who wish to explore the technical aspects of these analyses may go to www.metproject.org to find the three companion research reports: *Have We Identified Effective Teachers?* by Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger; *A Composite Estimator of Effective Teaching* by Kata Mihaly, Daniel F. McCaffrey, Douglas O. Staiger, and J.R. Lockwood; and *The Reliability of Classroom Observations by School Personnel* by Andrew D. Ho and Thomas J. Kane.

Earlier MET project briefs and research reports also on the website include:

Working with Teachers to Develop Fair and Reliable Measures of Teaching (2010). A white paper describing the rationale for and components of the MET project's study of multiple measures of effective teaching.

Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project (2010). A research report and non-technical policy brief with the same title on analysis of student-perception surveys and student achievement gain measures.

Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains (2012). A research report and policy/practitioner brief with the same title with initial findings on the reliability of classroom observations and implications for combining measures of teaching.

Asking Students about Teaching: Student Perception Surveys and Their Implementation (2012). A non-technical brief for policymakers and practitioners on the qualities of well-designed student surveys and implications for their implementation for teacher feedback and evaluation.

In addition, on www.metproject.org readers will find a set of principles to guide the design of teacher evaluation and support systems based on the work of the MET project, its partners, and other leading systems and organizations, *Feedback for Better Teaching: Nine Principles for Using Measures of Effective Teaching* (2013).

ABOUT THE MET PROJECT: The MET project is a research partnership of academics, teachers, and education organizations committed to investigating better ways to identify and develop effective teaching. Funding is provided by the Bill & Melinda Gates Foundation.

The approximately 3,000 MET project teachers who volunteered to open up their classrooms for this work are from the following districts: The Charlotte-Mecklenburg Schools, the Dallas Independent Schools, the Denver Public Schools, the Hillsborough County Public Schools, the Memphis Public Schools, the New York City Schools, and the Pittsburgh Public Schools.

Partners include representatives of the following institutions and organizations: American Institutes for Research, Cambridge Education, University of Chicago, The Danielson Group, Dartmouth College, Educational Testing Service, Empirical Education, Harvard University, National Board for Professional Teaching Standards, National Math and Science Initiative, New Teacher Center, University of Michigan, RAND, Rutgers University, University of Southern California, Stanford University, Teachscape, University of Texas, University of Virginia, University of Washington, and Westat.

MET Project Teachers

January 2013

Contents

Executive Summary 3

Can Measures of Effective Teaching Identify Teachers Who Better Help Students Learn? 6

How Much Weight Should Be Placed on Each Measure of Effective Teaching? 10

How Can Teachers Be Assured Trustworthy Results from Classroom Observations? 16

What We Know Now 20

Endnotes 23

Executive Summary

States and districts have launched unprecedented efforts in recent years to build new feedback and evaluation systems that support teacher growth and development. The goal is to improve practice so that teachers can better help their students graduate from high school ready to succeed in college and beyond.

These systems depend on trustworthy information about teaching effectiveness—information that recognizes the complexity of teaching and is trusted by both teachers and administrators. To that end, the Measures of Effective Teaching (MET) project set out three years ago to investigate how a set of measures could identify effective teaching fairly and reliably. With the help of 3,000 teacher volunteers who opened up their classrooms to us—along with scores of academic and organizational partners—we have studied, among other measures:

- **Classroom observation instruments,** including both subject-specific and cross-subject tools, that define discrete teaching competencies and describe different levels of performance for each;
- **Student perception surveys** that assess key characteristics of the classroom environment, including supportiveness, challenge, and order; and
- **Student achievement gains** on state tests and on more cognitively challenging assessments.

We have reported findings as we learned them in order to provide states and districts with evidence-based guidance to inform their ongoing work. In our initial report in 2010 (*Learning about Teaching*), we found that a well-designed student perception survey can provide reliable feedback on aspects of teaching practice that are predictive of student learning. In 2012 (*Gathering Feedback for Teaching*), we presented similar results for classroom observations. We also found that an accurate observation rating requires two or more lessons, each scored by a different certified observer. With each analysis we have better understood the particular contribution that each measure makes to a complete picture of effective teaching and how those measures should be implemented to provide teachers with accurate and meaningful feedback.

This final brief from the MET project's three-year study highlights new analyses that extend and deepen the insights from our previous work. These studies address three fundamental questions that face practitioners and policymakers engaged in creating teacher support and evaluation systems.

"Feedback and evaluation systems depend on trustworthy information about teaching effectiveness to support improvement in teachers' practice and better outcomes for students."

The Questions

Can measures of effective teaching identify teachers who better help students learn?

Despite decades of research suggesting that teachers are the most important in-school factor affecting student learning, an underlying question remains unanswered: Are seemingly more effective teachers truly better than other teachers at improving student learning, or do they simply have better students?

Ultimately, the only way to resolve that question was by randomly assigning students to teachers to see if teachers previously identified as more effective actually caused those students to learn more. That is what we did for a subset of MET project teachers. Based on data we collected during the 2009–10 school year, we produced estimates of teaching effectiveness for each teacher. We adjusted our estimates to account for student differences in prior test scores, demographics, and other traits. We then randomly assigned a classroom of students to each participating teacher for 2010–11.

Following the 2010–11 school year we asked two questions: First, did students actually learn more when randomly assigned to the teachers who seemed more effective when we evaluated them the prior year? And, second, did the magnitude of the difference in student outcomes following random assignment correspond with expectations?

How much weight should be placed on each measure of effective teaching?

While using multiple measures to provide feedback to teachers, many states and districts also are combining measures into a single index to support decisionmaking. To date, there has been little empirical evidence to inform how systems might weight each measure within a composite to support improvements in teacher effectiveness. To help fill that void, we tasked a group of our research partners to use data from MET project teachers to build and compare composites using different weights and different outcomes.

How can teachers be assured trustworthy results from classroom observations?

Our last report on classroom observations prompted numerous questions from practitioners about how to best use resources to produce quality information for feedback on classroom practice. For example: How many observers are needed to achieve sufficient reliability from a given number of observations? Do all observations need to be the same length to have confidence in the results? And what is the value of adding observers from outside a teacher's own school? To help answer these questions, we designed a study in which administrators and peer observers produced more than 3,000 scores for lessons taught by teachers within one MET project partner school district.

Key findings from those analyses:

1. **Effective teaching can be measured.** We collected measures of teaching during 2009–10. We adjusted those measures for the backgrounds and prior achievement of the students in each class. But, without random assignment, we had no way to know if the adjustments we made were sufficient to discern the markers of effective teaching from the unmeasured aspects of students' backgrounds.

In fact, we learned that the adjusted measures did identify teachers who produced higher (and lower) average student achievement gains following random assignment in 2010–11. The data show that we can identify groups of teachers who are more effective in helping students learn. Moreover, the magnitude of the achievement gains that teachers generated was consistent with expectations.

In addition, we found that more effective teachers not only caused students to perform better on state tests, but they also caused students to score higher on other, more cognitively challenging assessments in math and English.

2. **Balanced weights indicate multiple aspects of effective teaching.** A composite with weights between 33 percent and 50 percent assigned to state test scores demonstrated the best mix of low volatility from year to year and ability to predict student gains on multiple assessments. The composite that best indicated improvement on state tests heavily weighted teachers' prior student achievement gains based on those same tests. But composites that assigned 33 percent to 50 percent of the weight to state tests did nearly as well and were somewhat better at predicting student learning on more cognitively challenging assessments.

   Multiple measures also produce more consistent ratings than student achievement measures alone. Estimates of teachers' effectiveness are more stable from year to year when they combine classroom observations, student surveys, and measures of student achievement gains than when they are based solely on the latter.

3. **Adding a second observer increases reliability significantly more than having the same observer score an additional lesson.** Teachers' observation scores vary more from observer to observer than from lesson to lesson. Given the same total number of observations, including the perspectives of two or more observers per teacher greatly enhances reliability. Our study of video-based observation scoring also revealed that:

   a. Additional shorter observations can increase reliability. Our analysis suggests that having additional observers watch just part of a lesson may be a cost-effective way to boost reliability by including additional perspectives.

   b. Although school administrators rate their own teachers somewhat higher than do outside observers, how they rank their teachers' practice is very similar and teachers' own administrators actually discern bigger differences in teaching practice, which increases reliability.

   c. Adding observations by observers from outside a teacher's school to those carried out by a teacher's own administrator can provide an ongoing check against in-school bias. This could be done for a sample of teachers rather than all, as we said in *Gathering Feedback for Teaching.*

The following pages further explain these findings and the analyses that produced them.

Can Measures of Effective Teaching Identify Teachers Who Better Help Students Learn?[2]

By definition, teaching is effective when it enables student learning. But identifying effective teaching is complicated by the fact that teachers often have very different students. Students start the year with different achievement levels and different needs. Moreover, some teachers tend to get particular types of students year after year (that is, they tend to get higher-performing or lower-performing ones). This is why so-called value-added measures attempt to account for differences in the measurable characteristics of a teacher's students, such as prior test scores and poverty.

"Teachers previously identified as more effective caused students to learn more. Groups of teachers who had been identified as less effective caused students to learn less."

However, students differ in other ways—such as behavior and parental involvement—which we typically cannot account for in determining teaching effectiveness. If those "unaccounted for" differences also affect student learning, then what seems like effective teaching may actually reflect unmeasured characteristics of a teacher's students. The only way to know if measures of teaching truly identify effective teaching and not some unmeasured student characteristics is by randomly assigning teachers to students. So we did.

In 2009–10, we measured teachers' effectiveness using a combined measure, comprising teachers' classroom observation results, student perception survey responses, and student achievement gains adjusted for student characteristics, such as prior performance and demographics. The following year (2010–11), we randomly assigned different rosters of students to two or more MET project teachers who taught the same grade and subject in the same school. Principals created rosters and the RAND Corp assigned them randomly to teachers (see **Figure 1**). Our aim was to determine if the students who were randomly assigned to teachers who previously had been identified as more effective actually performed better at the end of the 2010–11 school year.[3]

They did. On average, the 2009–10 composite measure of effective teaching accurately predicted 2010–11 student performance. The research confirmed that, as a group, teachers previously identified as more effective caused students to learn more. Groups of teachers who had been identified as less effective

caused students to learn less. We can say they "caused" more (or less) student learning because when we randomly assigned teachers to students during the second year, we could be confident that any subsequent differences in achievement were being driven by the teachers, not by the unmeasured characteristics of their students. In addition, the magnitude of the gains they caused was consistent with our expectations.

Figure 2 illustrates just how well the measures of effective teaching predicted student achievement following random assignment. The diagonal line represents perfect prediction. Dots above the diagonal line indicate groups of teachers whose student outcomes following random assignment were better than predicted. Dots below the line indicate groups of teachers whose student outcomes following random assignment were worse than predicted. Each dot

Figure 1

Putting Measures of Effective Teaching to the Test with Random Assignment

1. Principals created rosters for each class
2. The rosters were assigned randomly within each grade and subject
3. We predicted student outcomes based on teachers' previous results, observations, and student surveys.
4. We compared those predictions to actual differences.

Do measures of teaching really identify teachers who help students learn more, or do seemingly more effective teachers just get better students? To find out, the MET project orchestrated a large-scale experiment with MET project teachers to see if teachers identified as more effective than their peers would have greater student achievement gains even with students who were assigned randomly.

To do so, the MET project first estimated teachers' effectiveness using multiple measures from the 2009–10 school year. As is common in schools, some teachers had been assigned students with stronger prior achievement than others. In assessing each teacher's practice that year, the project controlled for students' prior achievement and demographic characteristics. But there may have been other differences among students as well. So for the following school year (2010–11), principals created rosters of students for each class in the study, and then researchers randomly assigned each roster to a participating teacher from among those who could teach the class.

At the end of the 2010–11 school year, MET project analysts checked to see if students taught by teachers identified as more effective than their colleagues actually had greater achievement gains than students taught by teachers identified as less effective. They also checked to see how well actual student achievement gains for teachers matched predicted gains.

represents 5 percent of the teachers in the analysis, sorted based on their predicted impact on student achievement.[4]

As seen in **Figure 2**, in both math and English language arts (ELA), the groups of teachers with greater predicted impacts on student achievement generally had greater actual impacts on student achievement following random assignment. Further, the actual impacts are approximately in line with the predicted impacts.[5] We also found that teachers who we identified as being effective in promoting achievement on the state tests also generated larger gains on the supplemental tests administered in spring 2011.

Based on our analysis, we can unambiguously say that school systems should account for the prior test scores of students. When we removed this control, we wound up predicting much larger differences in achievement than actually occurred, indicating that student assignment biased the results. However, our analysis could not shed as much light on the need to control for demographics or "peer effects"—that is, the average prior achievement and demographics of each student's classmates. Although we included those

Figure 2

Effectiveness Measures Identify Teachers Who Help Students Learn More

Actual and Predicted Achievement of Randomized Classrooms (Math)

Actual and Predicted Achievement of Randomized Classrooms (English Language Arts)

These charts compare the actual 2010–11 school year achievement gains for randomly assigned classrooms with the results that were predicted based on the earlier measures of teaching effectiveness. Each dot represents the combination of actual and estimated student performance for 5 percent of the teachers in the study, grouped by the teachers' estimated effectiveness. The dashed line shows where the dots would be if the actual and predicted gains matched perfectly.

On average, students of teachers with higher teacher effectiveness estimates outperformed students of teachers with lower teacher effectiveness estimates. Moreover, the magnitude of students' actual gains largely corresponded with gains predicted by their effectiveness measured the previous year. Both the actual and predicted achievement are reported relative to the mean in the randomization block. That is, a zero on either axis implies that the value was no different from the mean for the small group of teachers in a grade, subject, and school within which class lists were randomized.

Impacts are reported in student-level standard deviations. A .25 standard deviation difference is roughly equivalent to a year of schooling. The predicted impacts are adjusted downward to account for incomplete compliance with randomization.

> "We can unambiguously say that school systems should adjust their achievement gain measures to account for the prior test scores of students. When we removed this control, we wound up predicting much larger differences in achievement than actually occurred."

controls, we cannot determine from our evidence whether school systems should include them. Our results were ambiguous on that score.

To avoid over-interpretation of these results, we hasten to add two caveats: First, a prediction can be correct on average but still be subject to measurement error. Our predictions of students' achievement following random assignment were correct on average, but within every group there were some teachers whose students performed better than predicted and some whose students performed worse. Second, we could not, as a practical matter, randomly assign students or teachers to a different school site. As a result, our study does not allow us to investigate bias in teacher effectiveness measures arising from student sorting between different schools.[6]

Nonetheless, our analysis should give heart to those who have invested considerable effort to develop practices and policies to measure and support effective teaching. Through this large-scale study involving random assignment of teachers to students, we are confident that we can identify groups of teachers who are comparatively more effective than their peers in helping students learn. Great teaching does make a difference.

How Much Weight Should Be Placed on Each Measure of Effective Teaching?[7]

Teaching is too complex for any single measure of performance to capture it accurately. Identifying great teachers requires multiple measures. While states and districts embrace multiple measures for targeted feedback, many also are combining measures into a single index, or composite. An index or composite can be a useful summary of complex information to support decisionmaking. The challenge is to combine measures in ways that support effective teaching while avoiding such unintended consequences as too-narrow a focus on one aspect of effective teaching.

To date, there has been little empirical evidence to suggest a rationale for particular weights. The MET project's report *Gathering Feedback for Teaching* showed that equally weighting three measures, including achievement gains, did a better job predicting teachers' success (across several student outcomes) than teachers' years of experience and masters' degrees. But that work did not attempt to determine optimal weights for composite measures.

Over the past year, a team of MET project researchers from the RAND Corporation and Dartmouth College used MET project data to compare differently weighted composites and study the implications of different weighting schemes for different outcomes. As in the *Gathering Feedback for Teaching* report, these composites included student achievement gains based on state assessments, classroom observations, and student surveys. The researchers estimated the ability of variously weighted composites to produce consistent results and accurately forecast teachers' impact on student achievement gains on different types of tests.

The goal was not to suggest a specific set of weights but to illustrate the trade-offs involved when choosing weights. Assigning significant weight to one measure might yield the best predictor of future performance on that measure. But heavily weighting a single measure may incentivize teachers to focus too narrowly on a single aspect

of effective teaching and neglect its other important aspects. For example, a singular focus on state tests could displace gains on other harder-to-measure outcomes. Moreover, if the goal is for students to meet a broader set of learning objectives than are measured by a state's tests, then too-heavily weighting that test could make it harder to identify teachers who are producing other valued outcomes.

Composites Compared

The research team compared four different weighting models, illustrated in **Figure 3:** (Model 1) The "best predictor" of state achievement test gains (with weights calculated to maximize the ability to predict teachers' student achievement gains on state tests, resulting in 65+ percent of the weight being placed on the student achievement gains across grades and subjects); (Model 2) a composite that assigned 50 percent of the weight to students' state achievement test gains; (Model 3) a composite that applied equal weights to each measure; and (Model 4) one that gave 50 percent to observation ratings and 25 percent each to achievement gains and student surveys. The weights that best predict state tests, shown for Model 1 in **Figure 3,** were calculated to predict gains on state ELA tests at the middle school level, which assigns a whopping 81 percent of the weight to prior gains on the same tests (best-predictor weights for other grades and subjects are in the table on page 14).

Figure 4 compares the different weighting schemes on three criteria, using middle school ELA as an example (see the table on page 14 for other grades and subjects). The first is predicting teachers' student achievement gains on state assessments. A correlation of 1.0 would indicate perfect accuracy in

> "Heavily weighting a single measure may incentivize teachers to focus too narrowly on a single aspect of effective teaching and neglect its other important aspects. ... [I]f the goal is for students to meet a broader set of learning objectives than are measured by a state's tests, then too-heavily weighting that test could make it harder to identify teachers who are producing other valued outcomes."

Figure 3

Four Ways to Weight

**Weights shown for Model 1 were calculated to best predict gains on state tests for middle school English language arts. Similar best predictor weights for other grades and subjects are in the table on page 14.*

These charts illustrate four ways to construct a composite measure of effective teaching. Each model uses different weights but includes the same components— student achievement gains on the state tests, student perception surveys, and classroom observations. Model 1 uses the weights that would best predict a teacher's impact on state test scores. Across grades and subjects, the "best predictor" model assigns 65 percent or more of the weight to a teacher's prior state test gains. Models 2–4 are not based on maximizing any particular outcome. They approximate different weighting schemes used by states and districts, with each model placing progressively less weight on student achievement gains on state tests.

Figure 4

Trade-Offs from Different Weighting Schemes

Middle School English Language Arts

These bars compare the four weighting schemes in Figure 3 on three criteria: accuracy in predicting teachers' achievement gains on state tests; accuracy in predicting student achievement gains on supplemental assessments designed to test higher-order thinking skills; and reliability, reflecting the year-to-year stability of teachers' results. Shown are the results for middle school ELA (see Table 1 on page 14 for results for other grades and subjects).

As indicated, Model 2 (50 percent state test results) and Model 3 (33 percent state tests) achieve much of the same predictive power as Model 1 (the "best predictor" of state test results) in anticipating teachers' future state test results (Model 1). Model 4 (50 percent observation) is considerably less predictive. However, the figures also illustrate two other trade-offs. Models 2 and 3 also are somewhat better than Model 1 at predicting gains on the tests of higher-order thinking skills (for all but elementary school math). Across most grades and subjects, Model 1 was the least reliable.

predicting teachers' student achievement gains on state tests. By definition, the best composite in this regard is Model 1, the model weighted for maximizing accuracy on state test results. Models 2–4 show the effect of reducing weights on student achievement gains on state tests for middle school ELA. As shown from middle school ELA, reducing weights on student achievement gains decreases the power to predict future student achievement gains on state tests from 0.69 to 0.63 with Model 2; to 0.53 with Model 3; and to 0.43 with Model 4. Other grades and subjects showed similar patterns, as indicated in the table on page 14.

While it is true that the state tests are limited and that schools should value other outcomes, observations and student surveys may not be more correlated with those other outcomes than the state tests. As a result, we set out to test the strength of each model's correlation with another set of test outcomes. The middle set of bars in **Figure 4** compares the four models (see Figure 3)—each using state test results to measure achievement gains—on how well they would predict teachers' student achievement gains on supplemental tests that were administered in MET project teachers' classrooms: The SAT 9 Open-Ended Reading Assessment (SAT 9 OE) and the Balanced Assessment in Mathematics (BAM).

While covering less material than state tests, the SAT 9 OE and BAM assessments include more cognitively challenging items that require writing, analysis, and application of concepts, and they are meant to assess higher-order thinking skills. Sample items released by the assessment consortia for the new Common Core State Standards assessments are more similar to the items on these supplemental tests than the ones on the state assessments. Shown in **Figure 4** is the effect of reducing the weight on state test gains in predicting gains on these other assessments, again for middle school ELA. For most grades and subjects, Model 2 and Model 3 (50 percent state test and equal weights for all three measures) best predicted teachers' student achievement gains on these supplemental assessments, with little difference between the two models. The one exception was elementary school math, where Model 1 (best predictor) was best.

The third set of bars in **Figure 4** compares composites on their reliability—that is, the extent to which the composite would produce consistent results for the same teachers from year to year (on a scale from 0–1.0, with

Increasing Accuracy, Reducing Mistakes

When high-stakes decisions must be made, can these measures support them? Undoubtedly, that question will be repeated in school board meetings and in faculty break rooms around the country in the coming years.

The answer is yes, not because the measures are perfect (they are not), but because the combined measure is better on virtually every dimension than the measures in use now. There is no way to avoid the stakes attached to every hiring, retention, and pay decision. And deciding not to make a change is, after all, a decision. No measure is perfect, but better information should support better decisions.

In our report *Gathering Feedback for Teaching,* we compared the equally weighted measure (Model 3 in Figures 3 and 4) to two indicators that are almost universally used for pay or retention decisions today: teaching experience and possession of a master's degree. On every student outcome—the state tests, supplemental tests, student's self-reported level of effort and enjoyment in class—the teachers who excelled on the composite measure had better outcomes than those with high levels of teaching experience or a master's degree.

In addition, many districts currently require classroom observations, but they do not include student surveys or achievement gains. We tested whether observations alone are enough. Even with four full classroom observations (two by one observer and two by another), conducted by observers trained and certified by the Educational Testing Service, the observation-only model performed far worse than any of our multiple measures composites. (The correlations comparable to those in Figure 5 would have been .14 and .25 with the state tests and test of higher-order skills.)

Still, it is fair to ask, what might be done to reduce error? Many steps have been discussed in this and other reports from the project:

- First, if any type of student data is to be used—either from tests or from student surveys—school systems should give teachers a chance to correct errors in their student rosters.
- Second, classroom observers should not only be trained on the instrument. They should first demonstrate their accuracy by scoring videos or observing a class with a master observer.
- Third, observations should be done by more than one observer. A principal's observation is not enough. To ensure reliability, it is important to involve at least one other observer, either from inside or outside the school.
- Fourth, if multiple years of data on student achievement gains, observations, and student surveys are available, they should be used. For novice teachers and for systems implementing teacher evaluations for the first time, there may be only a single year available. We have demonstrated that a single year contains information worth acting on. But the information would be even better if it included multiple years. When multiple years of data are available they should be averaged (although some systems may choose to weight recent years more heavily).

1.0 representing perfect consistency and no volatility). Again, results shown are for middle school ELA. Across all grades and subjects, the most reliable composites were either Models 2 (50 percent state test) or 3 (equal weights). For all but middle school math, the least reliable composite was Model 1 (best predictor). Model 4 (50 percent observations) was somewhat less reliable than Model 2 (equal weights) for all grades and subjects. Although not shown, student achievement gains on state tests by themselves are less stable than all of the composites, with one exception: Model 4 (50 percent observations) is slightly less stable than achievement gains alone for middle school math.

General Implications

The intent of this analysis was not to recommend an ideal set of weights to use in every circumstance. Rather, our goal was to describe the trade-offs among different approaches.[8]

If the goal is to predict gains on state tests, then the composites that put 65+ percent of the weight on the student achievement gains on those tests will generally show the greatest accuracy. However, reducing the weights on the state test achievement gain measures to 50 percent or 33 percent generates two positive trade-offs: it increases stability (lessens volatility from year to year) and it also increases somewhat the correlation with tests other than the state tests.

However, it is possible to go too far. Lowering the weight on state test achievement gains below 33 percent, and raising the weight on observations to 50 percent and including student surveys at 25 percent, is counterproductive. It not only lowers the

Table 1

CALCULATED WEIGHTS FOR MAXIMUM ACCURACY IN PREDICTING GAINS ON STATE TESTS

	English Language Arts			Math		
	State Tests	Observations	Student Surveys	State Tests	Observations	Student Surveys
Elementary	65%	9%	25%	85%	5%	11%
Middle	81%	2%	17%	91%	4%	5%

RELIABILITY AND ACCURACY OF DIFFERENT WEIGHTING SCHEMES

		English Language Arts				Math			
		Weighted for Max State Test Accuracy	50% State Test	Equal Weights	50% Observations	Weighted for Max State Test Accuracy	50% State Test	Equal Weights	50% Observations
Elementary	Reliability	0.42	0.46	0.50	0.49	0.52	0.57	0.57	0.55
	Correlation with state test	0.61	0.59	0.53	0.45	0.72	0.65	0.54	0.46
	Correlation with higher-order test	0.35	0.37	0.37	0.35	0.31	0.29	0.25	0.20
Middle	Reliability	0.51	0.66	0.76	0.75	0.86	0.88	0.88	0.83
	Correlation with state test	0.69	0.63	0.53	0.43	0.92	0.84	0.73	0.65
	Correlation with higher-order test	0.29	0.34	0.33	0.32	0.38	0.44	0.45	0.45

correlation with state achievement gains; it can also lower reliability and the correlation with other types of testing outcomes.

Ultimately, states, local education authorities, and other stakeholders need to decide how to weight the measures in a composite. Our data suggest that assigning 50 percent or 33 percent of the weight to state test results maintains considerable predictive power, increases reliability, and potentially avoids the unintended negative consequences from assigning too-heavy weights to a single measure. Removing too much weight from state tests, however, may not be a good idea, given the lower predictive power and reliability of Model 4 (25 percent state tests). In short, there is a range of reasonable weights for a composite of multiple measures.

Validity and Content Knowledge for Teaching

Teachers shouldn't be asked to expend effort to improve something that doesn't help them achieve better outcomes for their students. If a measure is to be included in formal evaluation, then it should be shown that teachers who perform better on that measure are generally more effective in improving student outcomes. This test for "validity" has been central to the MET project's analyses. Measures that have passed this test include high-quality classroom observations, well-designed student-perception surveys, and teachers' prior records of student achievement gains on state tests.

Over the past year, MET project researchers have investigated another type of measure, called the Content Knowledge for Teaching (CKT) tests. These are meant to assess teachers' understanding of how students acquire and understand subject-specific skills and concepts in math and ELA. Developed by the Educational Testing Service and researchers at the University of Michigan, these tests are among the newest measures of teaching included in the MET project's analyses. Mostly multiple choice, the questions ask how to best represent ideas to students, assess student understanding, and determine sources of students' confusion.

The CKT tests studied by the MET project did not pass our test for validity. MET project teachers who performed better on the CKT tests were not substantively more effective in improving student achievement on the outcomes we measured. This was true whether student achievement was measured using state tests or the supplemental assessments of higher-order thinking skills. For this reason, the MET project did not include CKT results within its composite measure of effective teaching.

These results, however, speak to the validity of the current measure still early in its development in predicting achievement gains on particular student assessments—not to the importance of content-specific pedagogical knowledge. CKT as a concept remains promising. The teachers with higher CKT scores did seem to have somewhat higher scores on two subject-based classroom observation instruments: the Mathematical Quality of Instruction (MQI) and the Protocol for Language Arts Teacher Observations (PLATO). Moreover, the MET project's last report suggested that some content-specific observation instruments were better than cross-subject ones in identifying teachers who were more effective in improving student achievement in ELA and math. Researchers will continue to develop measures for assessing teachers' content-specific teaching knowledge and validating them as states create new assessments aligned to the Common Core State Standards. When they have been shown to be substantively related to a teacher's students' achievement gains, these should be considered for inclusion as part of a composite measure of effective teaching.

How Can Teachers Be Assured Trustworthy Results from Classroom Observations?[9]

Classroom observations can be powerful tools for professional growth. But for observations to be of value, they must reliably reflect what teachers do throughout the year, as opposed to the subjective impressions of a particular observer or some unusual aspect of a particular lesson. Teachers need to know they are being observed by the right people, with the right skills, and a sufficient number of times to produce trustworthy results. Given this, the challenge for school systems is to make the best use of resources to provide teachers with high-quality feedback to improve their practice.

"For the same total number of observations, incorporating additional observers increases reliability."

The MET project's report *Gathering Feedback for Teaching* showed the importance of averaging together multiple observations from multiple observers to boost reliability. Reliability represents the extent to which results reflect consistent aspects of a teacher's practice, as opposed to other factors such as observer judgment. We also stressed that observers must be well-trained and assessed for accuracy before they score teachers' lessons.

But there were many practical questions the MET project couldn't answer in its previous study. Among them:

- Can school administrators reliably assess the practice of teachers in their schools?
- Can additional observations by external observers not familiar with a teacher increase reliability?
- Must all observations involve viewing the entire lesson or can partial lessons be used to increase reliability? And,
- What is the incremental benefit of adding additional lessons and additional observers?

These questions came from our partners, teachers, and administrators in urban school districts. In response, with the help of a partner district, the Hillsborough County (Fla.) Public Schools, the MET project added a study of classroom observation

Hillsborough County's Classroom Observation Instrument

Like many school districts, Hillsborough County uses an evaluation instrument adapted from the Framework for Teaching, developed by Charlotte Danielson. The framework defines four levels of performance for specific competencies in four domains of practice. Two of those domains pertain to activities outside the classroom: Planning and Preparation, and Professional Responsibility. Observers rated teachers on the 10 competencies in the framework's two classroom-focused domains, as shown:

Domain 2: The Classroom Environment	Domain 3: Instruction
▪ Creating an Environment of Respect and Rapport	▪ Communicating with Students
▪ Establishing a Culture of Learning	▪ Using Discussion and Questioning Techniques
▪ Managing Classroom Procedures	▪ Engaging Students in Learning
▪ Managing Student Behavior	▪ Using Assessment in Instruction
▪ Organizing Physical Space	▪ Demonstrating Flexibility and Responsiveness

reliability. This study engaged district administrators and teacher experts to observe video-recorded lessons of 67 Hillsborough County teachers who agreed to participate.

Comparison of Ratings

Two types of observers took part in the study: Fifty-three were school-based administrators—either principals or assistant principals—and 76 were peer observers. The latter are district-based positions filled by teachers on leave from the classroom who are responsible for observing and providing feedback to teachers in multiple schools. In Hillsborough County's evaluation system, teachers are observed multiple times, formally and informally, by their administrators and by peer observers. Administrators and peers are trained and certified in the district's observation instrument, which is based on Charlotte Danielson's Framework for Teaching. These observers each rated 24 lessons for us and produced more than 3,000 ratings that we could use to investigate our questions. MET project researchers were able to calculate reliability for many combinations of observers (administrator and peer), lessons (from 1 to 4), and observation duration (full lesson or 15 minutes). We were able to compare differences in the ratings given to teachers' lessons by their own and unknown administrators and between administrators and peers.

Effects on Reliability

Figure 5 graphically represents many of the key findings from our analyses of those ratings. Shown are the estimated reliabilities for results from a given set of classroom observations. Reliability is expressed on a scale from 0 to 1. A higher number indicates that results are more attributable to the particular teacher as opposed to other factors such as the particular observer or lesson. When results for the same teachers vary from lesson to lesson or from observer to observer, then averaging teachers' ratings across multiple lessons or observers decreases the amount of "error" due to such factors, and it increases reliability.

Adding lessons and observers increases the reliability of classroom observations. In our estimates, if a teacher's results are based on two lessons, having the second lesson scored by a second observer can boost reliability significantly. This is shown in **Figure 5:** When the same administrator observes a second lesson, reliability increases from .51 to .58, but when the second lesson is observed by a different administrator from the same school, reliability increases more than twice as much, from .51 to .67. Whenever a given number of lessons was split between multiple observers, the reliability was greater than that achieved by a single observer. In other words, for the same total number of observations, incorporating additional observers increases reliability.

Of course, it would be a problem if school administrators and peer observers produced vastly different results for the same teachers. But we didn't find that to be the case. Although administrators gave higher scores to their own teachers, their rankings of their own teachers were similar to those produced by peer observers and administrators from other schools. This implies that administrators are seeing the same

Figure 5

There Are Many Roads to Reliability

Lesson observed by own administrator = 45 min

Lesson observed by peer observer = 45 min

Three 15-minute lessons observed by three additional peer observers = 45 min

A and **B** denote different observers of the same type

These bars show how the number of observations and observers affects reliability. Reliability represents the extent to which the variation in results reflects consistent aspects of a teacher's practice, as opposed to other factors such as differing observer judgments. Different colors represent different categories of observers. The "A" and "B" in column three show that ratings were averaged from two different own-school observers. Each circle represents approximately 45 minutes of observation time (a solid circle indicates one observation of that duration, while a circle split into three indicates three 15-minute observations by three observers). As shown, reliabilities of .66–.72 can be achieved in multiple ways, with different combinations of number of observers and observations. (For example, one observation by a teacher's administrator when combined with three short, 15-minute observations each by a different observer would produce a reliability of .67.)

things in the videos that others do, and they are not being swayed by personal biases.

If additional observations by additional observers are important, how can the time for those added observations be divided up to maximize the use of limited resources while assuring trustworthy results? This is an increasingly relevant question as more school systems make use of video in providing teachers with feedback on their practice. Assuming multiple videos for a teacher exist, an observer could use the same amount of time to watch one full lesson or two or three partial lessons. But to consider the latter, one would want to know whether partial-lesson observations increase reliability.

Our analysis from Hillsborough County showed observations based on the first 15 minutes of lessons were about 60 percent as reliable as full lesson observations, while requiring one-third as much observer time. Therefore, one way to increase reliability is to expose a given teacher's practice to multiple perspectives. Having three different observers each observe for 15 minutes may be a more economical way to improve reliability than having one additional observer sit in for 45 minutes. Our results also suggest that it is important to have at least one or two full-length observations, given that some aspects of teaching scored on the Framework for Teaching (Danielson's instrument) were frequently not observed during the first 15 minutes of class.

> "Although administrators gave higher scores to their own teachers, their rankings of their own teachers were similar to those produced by external observers and administrators from other schools."

Together, these results provide a range of scenarios for achieving reliable classroom observations. There is a point where both additional observers and additional observations do little to reduce error. Reliability above 0.65 can be achieved with several configurations (see **Figure 5**).

Implications for Districts

Ultimately, districts must decide how to allocate time and resources to classroom observations. The answers to the questions of how many lessons, of what duration, and conducted by whom are informed by reliability considerations, as well as other relevant factors, such as novice teacher status, prior effectiveness ratings, and a district's overall professional development strategy.

What We Know Now

In three years we have learned a lot about how multiple measures can identify effective teaching and the contribution that teachers can make to student learning. The goal is for such measures to inform state and district efforts to support improvements in teaching to benefit all students. Many of these lessons have already been put into practice as school systems eagerly seek out evidence-based guidance. Only a few years ago the norm for teacher evaluation was to assign "satisfactory" ratings to nearly all teachers evaluated while providing virtually no useful information to improve practice.[10] Among the significant lessons learned through the MET project and the work of its partners:

- **Student perception surveys and classroom observations can provide meaningful feedback to teachers.** They also can help system leaders prioritize their investments in professional development to target the biggest gaps between teachers' actual practice and the expectations for effective teaching.
- **Implementing specific procedures in evaluation systems can increase trust in the data and the results.** These include rigorous training and certification of observers; observation of multiple lessons by different observers; and in the case of student surveys, the assurance of student confidentiality.
- **Each measure adds something of value.** Classroom observations provide rich feedback on practice. Student perception surveys provide a reliable indicator of the learning environment and give voice to the intended beneficiaries of instruction. Student learning gains (adjusted to account for differences among students) can help identify groups of teachers who, by virtue of their instruction, are helping students learn more.
- **A balanced approach is most sensible when assigning weights to form a composite measure.** Compared with schemes that heavily weight one measure, those that assign 33 percent to 50 percent of the weight to student achievement gains achieve more consistency, avoid the risk of encouraging too narrow a focus on any one aspect of teaching, and can support a broader range of learning objectives than measured by a single test.
- **There is great potential in using video for teacher feedback and for the training and assessment of observers.** The advances made in this technology have been significant, resulting in lower costs, greater ease of use, and better quality.

The Work Ahead

As we move forward, MET project teachers are supporting the transition from research to practice. More than 300 teachers are helping the project build a video library of practice for use in professional development. They will record more than 50 lessons each by the end of this school year and make these lessons available to states, school districts, and other organizations committed to improving effective teaching. This will allow countless educators to analyze instruction and see examples of great teaching in action.

Furthermore, the unprecedented data collected by the MET project over the past three years are being made available to the larger research community to carry out additional analyses, which will increase knowledge of what constitutes effective teaching and how to support it. MET project partners already are tapping those data for new studies on observer training, combining student surveys and observations, and other practical concerns. Finally, commercially available video-based tools for observer training and certification now exist using the lessons learned from the MET project's studies.

Many of the future lessons regarding teacher feedback and evaluation systems must necessarily come from the field, as states and districts innovate, assess the results, and make needed adjustments. This will be a significant undertaking, as systems work to better support great teaching. Thanks to the hard work of MET project partners, we have a solid foundation on which to build.

"Many of the future lessons regarding teacher feedback and evaluation systems must necessarily come from the field, as states and districts innovate, assess the results, and make needed adjustments. This will be a significant undertaking, as systems work to better support great teaching."

Endnotes

1. The lead authors of this brief are Steven Cantrell, Chief Research Officer at the Bill & Melinda Gates Foundation, and Thomas J. Kane, Professor of Education and Economics at the Harvard Graduate School of Education and principal investigator of the Measures of Effective Teaching (MET) project. Lead authors of the related research papers are Thomas J. Kane (Harvard), Daniel F. McCaffrey (RAND), and Douglas O. Staiger (Dartmouth). Essential support came from Jeff Archer, Sarah Buhayar, Alejandro Ganimian, Andrew Ho, Kerri Kerr, Erin McGoldrick, and David Parker. KSA-Plus Communications provided design and editorial assistance.

2. This section summarizes the analyses and key findings from the research report *Have We Identified Effective Teachers?* by Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. Readers who want to review the full set of findings can download that report at www.metproject.org.

3. As expected, not every student on a randomly assigned roster stayed in the classroom of the intended teacher. Fortunately, we could track those students. We estimated the effects of teachers on student achievement using a statistical technique commonly used in randomized trials called "instrumental variables."

4. These predictions, as well as the average achievement outcomes, are reported relative to the average among participating teachers in the same school, grade, and subject.

5. Readers may notice that some of the differences in Figure 2 are smaller than the differences reported in earlier MET reports. Due to non-compliance—students not remaining with their randomly assigned teacher—only about 30 percent of the randomly assigned difference in teacher effectiveness translated into differences in the effectiveness of students' actual teacher. The estimates in Figure 2 are adjusted for non-compliance. If all the students had remained with their randomly assigned teachers, we would have predicted impacts roughly three times as big. Our results imply that, without non-compliance, we would have expected to see differences just as large as included in earlier reports.

6. Other researchers have studied natural movements of teachers between schools (as opposed to randomly assigned transfers) and found no evidence of bias in estimated teacher effectiveness between schools. See Raj Chetty, John Friedman, and Jonah E. Rockoff, "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," working paper no. 17699, National Bureau of Economic Research, December 2011.

7. The findings highlighted in this summary and the technical details of the methods that produced them are explained in detail in the research paper "A Composite Estimator of Effective Teaching," by Kata Mihaly, Daniel McCaffrey, Douglas O. Staiger, and J.R. Lockwood. A copy may be found at www.metproject.org.

8. Different student assessments, observation protocols, and student survey instruments would likely yield somewhat different amounts of reliability and accuracy. Moreover, measures used for evaluation may produce different results than seen in the MET project, which attached no stakes to the measures it administered in the classrooms of its volunteer teachers.

9. This section summarizes key analyses and findings from the report *The Reliability of Classroom Observations by School Personnel* by Andrew D. Ho and Thomas J. Kane. Readers who want to review the full set of findings and methods for the analyses can download that report at www.metproject.org. The MET project acknowledges the hard work of Danni Greenberg Resnick and David Steele, of the Hillsborough County Public Schools, and the work of the teachers, administrators, and peer observers who participated in this study.

10. Weisburg, D. et al. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness.* Brooklyn: New Teacher Project.

Bill & Melinda Gates Foundation

Guided by the belief that every life has equal value, the Bill & Melinda Gates Foundation works to help all people lead healthy, productive lives. In developing countries, it focuses on improving people's health and giving them the chance to lift themselves out of hunger and extreme poverty. In the United States, it seeks to ensure that all people—especially those with the fewest resources—have access to the opportunities they need to succeed in school and life. Based in Seattle, Washington, the foundation is led by CEO Jeff Raikes and Co-chair William H. Gates Sr., under the direction of Bill and Melinda Gates and Warren Buffett.

For more information on the U.S. Program, which works primarily to improve high school and postsecondary education, please visit www.gatesfoundation.org.

©2013 Bill & Melinda Gates Foundation. All Rights Reserved.
Bill & Melinda Gates Foundation is a registered trademark in the United States and other countries.

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org

EXHIBIT 3
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

Valuing Teachers

How much is a good teacher worth?

By ERIC A. HANUSHEK

For some time, we have recognized that the academic achievement of schoolchildren in this country threatens, to borrow President Barack Obama's words, "the U.S.'s role as an engine of scientific discovery" and ultimately its success in the global economy. The low achievement of American students, as reflected in the Program for International Student Assessment (PISA) (see "Teaching Math to the Talented," *features*, Winter 2011), will prevent them from accessing good, high-paying jobs. And, as demonstrated in another article in *Education Next* (see "Education and Economic Growth," *research*, Spring 2008), lower achievement means slower growth in the economy. From studying the historical relationship, we can estimate that closing just half of the performance gap with Finland, one of the top international performers in terms of student achievement, could add more than $50 trillion to our gross domestic product between 2010 and 2090. By way of comparison, the drop in economic output over the course of the last recession is believed to be less than $3 trillion. Thus the achievement gap between the U.S. and the world's top-performing countries can be said to be causing the equivalent of a permanent recession.

According to the president in this year's State of the Union address, this is "our generation's *Sputnik* moment," the time when we realize the urgent need to step up the performance of our education system. Only today, unlike in the 1950s, we have a clear idea of what it takes to improve achievement. The quality of the teachers in our schools is paramount: no other measured aspect of schools is nearly as important in determining student achievement. The initiatives we have emphasized in policy discussions—class-size reduction, curriculum revamping, reorganization of school schedule, investment in technology—all fall far short of the impact that good teachers can have in the classroom. Moreover, many of these interventions can be very costly.

Indeed, the magnitude of variation in the quality of teachers, even within each school, is startling. Teachers who work in a given school, and therefore teach students with similar demographic characteristics, can be responsible for increases in math and reading levels that range from a low of one-half year to a high of one and a half years of learning each academic year.

But while most parents are able to distinguish a good teacher from a bad one, few have any idea what difference it makes in the lives of their children. And researchers do not help, tending to talk in terms of standard deviations of achievement and effect sizes, phrases that simply have no meaning outside of the rarefied world of research. Here, I translate the researchers' shorthand into concepts that might be more readily understood: the impact of teachers on the earnings of individuals and on the future of the economy as a whole.

ILLUSTRATION / IMAGEZOO / GETTY IMAGES

Measuring Teachers' Impact

Many of us have had at some point in our lives a wonderful teacher, one whose value, in retrospect, seems inestimable. We do not pretend here to know how to calculate the life-transforming effects that such teachers can have with particular students. But we can calculate more prosaic economic values related to effective teaching, by drawing on a research literature that provides surprisingly precise estimates of the impact of student achievement levels on their lifetime earnings and by combining this with estimated impacts of more-effective teachers on student achievement.

Let's start with the researcher's point of view. With a normal distribution of performance (the classic bell curve), a standard deviation is simply a more precise measure of how spread out the distribution is. Somebody who is one standard deviation above average would be at the 84th percentile of the distribution. If we then turn to the labor market, a student with achievement (as measured by test performance in high school) that is one standard deviation above average can later in life expect to take in 10 to 15 percent higher earnings per year.

That estimate may be deemed conservative for two reasons. First, it does not account for increases in years of education that may result from having a higher level of performance early on. Also, the estimate is based on information from people's wages and salaries early in their careers, before they have reached their full earnings potential. Other calculations that take into account earnings throughout entire careers estimate 20 percent increases over the course of a lifetime.

A good, but not great, teacher increases each student's lifetime earnings by $10,600. Given a class of 20 students, she will raise their aggregate earnings by $212,000.

Does 10 to 15 percent amount to much? For the average American entering the labor force, the value of lifetime earnings for full-time work is currently $1.16 million. Thus, an increase in the level of achievement in high school of a standard deviation yields an average increase of between $110,000 and $230,000 in lifetime earnings.

How do increases in teacher effectiveness relate to this? Obviously, teacher quality is not the only factor that affects student achievement. The student's own motivations and support from family and peers play crucial roles as well. But researchers have worked hard to isolate the impact of teachers from these other influences. Rigorous studies consistently show that the impact of a more-effective teacher is substantial A high-performing teacher, one at the 84th percentile of all teachers, when compared with just an average teacher, produces students whose level of achievement is at least 0.2 standard deviations higher by the end of the school year. In fact, the impact of having such a teacher could plausibly be as large as 0.3 standard deviations.

Those impacts attenuate somewhat over time, however. The literature, though less than definitive, suggests that perhaps 70 percent of the gains achieved that year are retained in the long run by the student. The persistence of achievement gains is important, because the more sustained that these increases are, the greater the positive impact teachers will have on the lifetime skills and therefore the earnings of students. Put together, this evidence suggests that a teacher in the top 16 percent of effectiveness will have a positive impact (as compared to an average teacher) on longer-term student achievement that is 70 percent of the immediate gain, which as noted is at least 0.2 standard deviations. That lower bound of the estimated effect is what we will use as we calculate the economic worth of a teacher by combining a teacher's impact on achievement with the associated labor market returns.

Let's start with some conservative estimates of the impact on an individual student. Take a good but not great teacher, one at the 69th percentile of all teachers rather than at the 50th percentile (that is, a teacher who is half a standard deviation above the average). She produces an increase of $10,600 on each student's lifetime earnings. Even a modestly better than average teacher (60th percentile) raises individual earnings by $5,300, compared to what would otherwise be expected.

While those numbers are not trivial, they burgeon dramatically once we recognize that every student in the class can expect such increases in earnings. Consider, for example, a teacher with a class of 20 students. Under such circumstances, the teacher at the 60th percentile will—each year—raise students' aggregate earnings by a total of $106,000. The impact of one at the 69th percentile (as compared to the average) is $212,000, and one at the 84th percentile will shift earnings up by more than $400,000.

But there is also symmetry to these calculations. A very low performing teacher (at the 16th percentile of effectiveness) will have a negative impact of $400,000 compared to an average teacher.

Moreover, the economic value of an effective teacher grows with larger classes, as do the economic losses of an ineffective teacher. Figure 1 illustrates the aggregate impact on students'

lifetime earnings for higher- and lower-performing teachers. As we will discuss below, these results are all very large compared with, for instance, the $52,000 annual salary U.S. teachers were paid on average in 2008.

Effective Teachers Raise Students' Earnings
(Figure 1)

The economic value of an effective teacher grows with larger classes, and the economic costs of having an ineffective teacher are substantial.

SOURCE: Authors' calculations

An Alternate Thought Experiment

We can also approach this valuation calculation from the perspective of the impact of teacher effectiveness on the U.S. economy as a whole, rather than just on the future earnings of students. As noted above, student achievement, which provides a direct measure of later quality of the labor force, is strongly related to economic growth. Improving achievement leads to a better prepared workforce and to greater growth, and this growth translates into higher levels of national income.

Starting again with the estimates of the difference in effectiveness of teachers, it is possible to calculate the long-term economic impact of policies that would focus attention on the lowest-quality teachers from U.S. classrooms. Let us propose the following thought experiment: What would happen if the very lowest performing teachers could be replaced by just average teachers? Based on the estimates of variation in teacher quality identified above, Figure 2 shows the overall achievement impact through a cycle of K–12 instruction. Assuming the upper-bound estimate of teachers' impact, U.S. achievement could reach that in Canada and Finland if we replaced with average teachers the least effective 5 to 7 percent of teachers, respectively. Assuming the lower-bound estimate of teachers' impact, U.S. achievement could reach that in Canada and Finland if we replaced with average teachers the least effective 8 to 12 percent of teachers, respectively.

Here the estimated value almost loses any meaning. Closing the achievement gap with Finland would, according to historical experience, have astounding benefits, increasing the annual growth rate of the United States by 1 percent of GDP. Accumulated over the lifetime of somebody born today, this improvement in achievement would amount to nothing less than an increase in total U.S. economic output of $112 trillion in present value. (That was not a typo—$112 *trillion*, not billion.)

Admittedly, these estimates are subject to some uncertainty. So if you think those that are given here are too high, even though they are based on the best of contemporary research, then just cut them in half. You will still have effects on growth of one-half of 1 percent per year, which produces impacts of $56 trillion over the lifetime of today's child. In other words, to make the very large effects disappear, you have to make either the very strong assumption that student learning has little effect on the U.S. economy or the equally strong assumption that teachers have little impact on students.

What Would It Take?

The majority of our teachers are hardworking and effective. The previous estimates point clearly to the key imperative of eliminating the drag of the bottom teachers. Here we can offer several alternatives.

One approach might be better recruitment so that ineffective or poor teachers do not make it into our schools. Or, relatedly, we could improve the training in schools of education so that the average teaching recruit is better than the typical recruit of today. Unfortunately, we have

relatively few successful experiences with either approach as compared to considerable wishful thinking, particularly among school personnel.

An alternative might be to change a poor teacher into an average teacher. This approach is in fact today's dominant strategy. Schools hope that through mentoring of incoming teachers, professional development, or completion of further graduate schooling, ineffective teachers can be transformed into acceptable (average) teachers. Again, however, the existing evidence is not very reassuring. While such efforts undoubtedly help some teachers, there is no substantial evidence that certification, in-service training, master's degrees, or mentoring programs systematically make a difference in whether teachers are in fact effective at driving student achievement.

Measuring Up (Figure 2)

The U.S. could reach the achievement levels attained by such countries as Canada and Finland by replacing the lowest-performing teachers with average teachers.

Note: As derived from studies of teacher effectiveness, the lower bound assumes that a teacher at the 16th percentile of the distribution will obtain learning gains that are 0.2 standard deviations less than the average teacher obtains. The upper bound corresponds to 0.3 standard deviations less.

SOURCE: Authors' calculations

The final option is a clearer evaluation and retention strategy for teachers. Today, obtaining an entry job into teaching is virtually tantamount to an indefinite contract that stays in force regardless of actual effectiveness in the classroom. Yet the calculations above show the enormous value to individuals and society of "deselecting" the least effective teachers.

Is such a policy change feasible? If we contemplate asking 5 to 10 percent of teachers to find a job at which they are more effective so they can be replaced by teachers of average productivity, states and school districts would have to change their employment practices. They would need recruitment, pay, and retention policies that allow for the identification and compensation of teachers on the basis of their effectiveness with students. At a minimum, the current dysfunctional teacher-evaluation systems would need to be overhauled so that effectiveness in the classroom is clearly identified. This is not an impossible task. The teachers who are excellent would have to be paid much more, both to compensate for the new riskiness of the profession and to increase the chances of retaining these individuals in teaching. Those who are ineffective would have to be identified and replaced. Both steps would be politically challenging in a heavily unionized environment such as the one in place today.

Salary Politics

The above discussion also highlights the difficulties in recruiting high-quality teachers, due in part to the difficulties of paying them well. Collective bargaining mechanisms do not provide incentives for the best people to enter or remain in the profession and likely hold the average pay down: given the uniform salary structure, increases in salary are bound to be unrelated to increases in effectiveness, making large pay raises politically problematic. This is likely one of the main reasons that teacher salaries now lag those in other professions. In the 1940s, the salaries of male teachers were slightly above the average pay for all male college graduates, and female teachers had higher salaries than 70 percent of other female college graduates. Today, despite the collective bargaining process, the salaries of male teachers are at the 30th percentile of the distribution of all college graduates, and women who teach are at the 40th percentile of their college-educated peers.

Teachers' salaries today are based on credentials and years of experience, factors that are at best weakly related

Unless we can replace the current system with one that better links teacher recruitment, compensation, and retention to effectiveness, we should expect both our schools and our economy to underperform relative to their potential.

to productivity. In a competitive marketplace, a firm must compensate employees according to their productivity or risk bankruptcy. Yet no school district goes out of business if it retains ineffective teachers and pays them as much as effective ones. Salaries become political footballs, and it is often awkward for politicians to explain why a large pay increase goes equally to ineffective and effective teachers.

The challenge of implementing reform of the teaching profession remains considerable. Most of the benefits of implementing the "thought experiment" explored here would be fully realized only many decades later, while the costs of economic, and especially political, reform must be paid at the beginning. These costs would be steep, as they would likely negatively affect some of the most vocal constituents in education policy: current teachers.

The magnitude of the above valuations of teacher effectiveness, however, suggest that we should be willing to consider more radical reforms than have been commonplace in recent decades. Salaries several times higher than those paid teachers today would be economically justified if teachers were compensated according to their effectiveness. But unless we can replace the current system with one that better links teacher recruitment, compensation, and retention to effectiveness, we should expect both our schools and our economy to underperform relative to their potential. The cost to the nation at a time of intensifying international competition is high indeed.

Eric A. Hanushek is a senior fellow at the Hoover Institution, Stanford University.

Join EWA today for free!
Benefits for active reporter & associate members include:
• Separate e-mail communities for K-12, higher-ed, & associate members to share ideas & resources
• Discounted seminars and free regional workshops
• Full access to the EWA members-only website
• Guidebooks & tips to covering issues
• Electronic newsletter, the Education Reporter
•Source lists, backgrounders, and other EWA publications free of charge

Contact Stephanie Cvetetic, Marketing and Development Coordinator, for sponsorship or membership information, scvetetic@ewa.org

NATIONAL
EWA
Education Writers Association

The **Education Writers Association** (EWA) is the national professional organization of education journalists and others interested in how education is covered in the media. EWA is dedicated to improving the quality and quantity of education coverage to create a more informed public. For more than 60 years, EWA has provided information, training, and support to education journalists. Today, EWA has more than 2,000 members directly benefiting from its services, staff, and understanding of the issues.

www.ewa.org

EXHIBIT 4
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

Econometrica, Vol. 73, No. 2 (March, 2005), 417–458

TEACHERS, SCHOOLS, AND ACADEMIC ACHIEVEMENT

BY STEVEN G. RIVKIN, ERIC A. HANUSHEK, AND JOHN F. KAIN[1]

This paper disentangles the impact of schools and teachers in influencing achievement with special attention given to the potential problems of omitted or mismeasured variables and of student and school selection. Unique matched panel data from the UTD Texas Schools Project permit the identification of teacher quality based on student performance along with the impact of specific, measured components of teachers and schools. Semiparametric lower bound estimates of the variance in teacher quality based entirely on within-school heterogeneity indicate that teachers have powerful effects on reading and mathematics achievement, though little of the variation in teacher quality is explained by observable characteristics such as education or experience. The results suggest that the effects of a costly ten student reduction in class size are smaller than the benefit of moving one standard deviation up the teacher quality distribution, highlighting the importance of teacher effectiveness in the determination of school quality.

KEYWORDS: Student achievement, teacher quality, school selection, class size, teacher experience.

1. INTRODUCTION

SINCE THE RELEASE of *Equality of Educational Opportunity* (the "Coleman Report") in 1966, the educational policy debate in the United States and elsewhere has often been reduced to a series of simplistic arguments and assertions about the role of schools in producing achievement.[2] The character of this debate has itself been heavily influenced by confusing and conflicting research. While this research has frequently suffered from inadequate data, imprecise formulation of the underlying problems and issues has been as important in obscuring the fundamental policy choices. This paper defines a series of basic issues about the performance of schools that are relevant for current policy debates and considers how observed student performance can be used to address

[1]While John Kain participated fully in this project, he sadly died before its publication. We are grateful to Kraig Singleton, Jaison George, and Dan O'Brien for excellent research assistance, and we thank Eric French, Caroline Hoxby, Jessica Wolpaw Reyes, Finis Welch, Geoffrey Woglom, and a co-editor, along with seminar participants at UC Berkeley, UC Davis, UC San Diego, the National Bureau of Economic Research, the Public Policy Research Institute, Stanford University, University of Texas, and Texas A&M University for their many helpful comments. The arguments and estimation were considerably strengthened by the comments of anonymous referees. Hanushek and Rivkin thank the Donner Foundation, the Smith Richardson Foundation, and the Packard Humanities Institute for funding, and Kain thanks the Smith-Richardson Foundation and the Spencer Foundation.

[2]The original Coleman Report (Coleman et al. (1966)) was subjected to considerable criticism both for methodology and interpretation; see, for example, Hanushek and Kain (1972). The ensuing controversy led to considerable new research, but this new work has not ended the controversy; see Hanushek (1996, 2003) and Greenwald, Hedges, and Laine (1996). Those discussions represent the starting point for this research.

each. It then employs a unique panel data set of students in Texas to identify the sources of differences in student achievement and the relevance of a broad class of policies related to school resources.

Some very basic questions that have arisen from prior work command a central position in most policy discussions. First, partly resulting from common misinterpretations of the Coleman Report, do schools "make a difference" or not? While a surprising amount of controversy continues over this issue, it comes down to a simple question of whether or not there are significant and systematic differences between schools and teachers in their abilities to raise achievement. Second, how important are any differences in teacher quality in the determination of student outcomes? Finally, are any quality differences captured by observable characteristics of teachers and schools including class size, teacher education, and teacher experience? If so, how large are the effects? This third issue is in fact the genesis of the first, because the Coleman Report reported relatively small effects of differences in the measured attributes of schools on student achievement—a finding that has frequently been interpreted as indicating that there are no systematic quality differences among schools.

An extraordinarily rich data set providing longitudinal information on individual achievement of students in the State of Texas permits analyses that yield quite precise answers to each of these questions. The data contain test scores spanning grades 3 through 7 for three cohorts of students in the mid-1990s. The multiple cohorts and grades coming from repeated observations on more than one-half million students in over three thousand schools permit the clear identification and detection of even very small teacher and school effects.

A primary objective of the initial empirical analysis is to obtain estimates of differences in teacher contributions to student learning that eliminate the major sources of possible contamination from student selection or teacher assignment practices. Because family choice of neighborhood and school depends on preferences and resources, students are nonrandomly distributed across schools (Tiebout (1956)). Schools also use student characteristics including assessments of ability and achievement to place students into specific programs and classes. Such nonrandom selection may easily contaminate estimates of school or teacher effects with the influences of unmeasured individual, family, school, and neighborhood factors.

Repeated performance observations for individual students and multiple cohorts provide a means of controlling explicitly for student heterogeneity and the nonrandom matching of students, teachers, and schools through the use of fixed effects models. The models control for fixed student, school-by-grade, and in some cases school-by-year effects and then relate remaining differences in achievement gains between grades and cohorts to differences in school characteristics or teachers. This variation in academic performance cannot be driven by unchanging student attributes such as ability or motivation or by unchanging school characteristics and policies that are either common across all

grades at a point in time or unique to specific grades. Moreover, the empirical models also account for potentially important time varying influences not captured by the student or school fixed effects. Therefore we are able to identify the impacts of schools and teachers uncontaminated by the many unobserved family and other influences that have plagued past research.

The results reveal large differences among teachers in their impacts on achievement and show that high quality instruction throughout primary school could substantially offset disadvantages associated with low socioeconomic background. These differences among teachers are not, however, readily measured by simple characteristics of the teachers and classrooms. Consistent with prior findings, there is no evidence that a master's degree raises teacher effectiveness. In addition, experience is not significantly related to achievement following the initial years in the profession. These findings explain much of the contradiction between the perceived role of teachers as the key determinant of school quality and the body of research showing that observed teacher characteristics including experience and education explain little of the variation in student achievement.

Students also appear to benefit from smaller classes, particularly in grades 4 and 5. In comparison to the gains from higher teacher quality, however, the estimates indicate that even a very costly ten student reduction in class size such as that undertaken in some U.S. states produces smaller benefits than a one standard deviation improvement in teacher quality.

The next section provides an overview of patterns of achievement gains that suggests the presence of substantial within school variation in teacher quality. Section 3 describes the empirical approach used to generate a lower bound estimate of the within school variation in teacher quality. Section 4 provides a detailed description of the Texas data on students and teachers. Section 5 reports estimates of the variance in teacher quality based on the method developed in Section 3, and Section 6 presents an extension of traditional analyses of the effects of measured resources: class size, teacher education, and teacher experience on achievement. The final section considers the policy implications of the findings, particularly the importance of measured resources relative to the overall contribution of teachers.

2. SCHOOLS AND TEACHERS

Students and parents refer often to differences in teacher quality and act to ensure placement in classes with specific teachers. Such emphasis on teachers is largely at odds with empirical research into teacher quality. There has been no consensus on the importance of specific teacher factors, leading to the common conclusion that the existing empirical evidence does not find a strong role for teachers in the determination of academic achievement and future academic and labor market success. It may be that parents and students overstate the importance of teachers, but an alternative explanation is that measurable

characteristics such as teacher experience, education, and even test scores of teachers explain little of the true variation in quality.

To motivate the concentration on teacher quality, we begin with aggregate statistics on the variation in student achievement. Table I displays correlations of school average annual mathematics and reading achievement gains in grades 5, 6, and 7 between two cohorts of students for all public elementary schools in Texas.[3] The diagonal elements report correlations for the same grade in adjacent years, while the off-diagonal elements report correlations for adjacent grades in the same year.

The striking difference in magnitudes of the diagonal and off-diagonal elements suggests the existence of substantial within-school heterogeneity in school quality. Remarkably, the correlation of between-cohort average gains in different grades in the same year (the off-diagonal terms) is quite small despite the homogeneity of family backgrounds and peers within most schools and despite the common school organization, leadership, and resources for the two cohorts. Indeed for comparisons of 6th and 7th grade reading performance, the correlation is −0.01. In contrast, the correlations of between-cohort average gains in the same grade in adjacent years (the diagonal terms) are much larger. A number of factors may explain this pattern, but perhaps the most obvious explanation is that there will be many common teachers for two cohorts when observed in the same grade, while virtually all of the teachers will be different when comparing cohort performance across grades at a single point in time.

Table II reports the R^2 values from a series of achievement gain regressions for reading and mathematics performance run over the sample of schools and grades in which there is only a single teacher per subject. (As we discuss be-

TABLE I

PEARSON CORRELATION COEFFICIENTS OF SCHOOL AVERAGE TEST SCORE GAINS IN MATHEMATICS AND READING ACROSS GRADES AND YEARS

Grade of Cohort I	Mathematics Grade of Cohort II			Reading Grade of Cohort II		
	5	6	7	5	6	7
5	0.32**			0.19**		
6	0.13**	0.52**		0.13**	0.43**	
7		0.05	0.46**		−0.01	0.44**

Notes: Cohort I attended 4th grade in 1994; Cohort II attended 4th grade in 1995. Thus, for example, Cohort I is attending the 6th grade during the same academic year that Cohort II is attending the 5th grade. All calculations are weighted by the average enrollment of the pairs.

*significant at 10% level; **significant at 1% level.

[3]These data, subsequently used in the detailed empirical analyses, are described in detail in Section 3. All correlations relate just to students in schools that have both of the relevant grades.

TABLE II

COMPARISON OF THE EXPLANATORY POWER OF TEACHER EXPERIENCE, EDUCATION, AND CLASS SIZE WITH TEACHER FIXED EFFECTS IN EXPLAINING ACHIEVEMENT GAINS

	Mathematics				Reading			
Included explanatory variables								
Student covariates	yes	yes	yes	yes	yes	yes	yes	yes
Teacher characteristics	no	yes	no	no	no	yes	no	no
Teacher fixed effects	no	no	yes	no	no	no	yes	no
School fixed effects	no	no	no	yes	no	no	no	yes
R squared	0.0151	0.0182	0.1640	0.0949	0.0085	0.0093	0.0903	0.0507
Observations	89,414				81,897			

Notes: Dependent variables are mathematics and reading test score gains; sample includes only grades in a school with a single teacher for that subject.

low, these are the only schools in which students can be matched to their actual teachers.) The first column for each subject is based on a specification with only student characteristics and year dummies; the second column adds measured teacher and classroom characteristics (teacher experience, teacher education, and class size); the third column substitutes teacher fixed effects for the observable teacher and classroom characteristics; and the final column employs school rather than teacher fixed effects. The results demonstrate quite clearly that the observable school and teacher characteristics explain little of the between-classroom variation in achievement growth despite the fact that a substantial share of the overall achievement gain variation occurs between teachers. Importantly, even though the sample includes just schools with a single teacher per grade, the inclusion of school rather than teacher fixed effects reduces the explanatory power by over forty percent, suggesting that much of the variation in teacher quality exists within rather than between schools.

Tables I and II are consistent with the existence of substantial variation in teacher quality not explained by observable teacher characteristics. However, other factors could clearly enter into these two simple comparisons, making it necessary to utilize more comprehensive methods to identify the variance of teacher quality and importance of observable factors. For example, a high performing 4th grade teacher could leave less room for subsequent gains; the curriculum could affect specific grade levels in differing ways across school districts; test measurement errors could obscure the relationships; there may be nonrandom sorting across schools; or some schools may have more or less effective leadership. The next section develops a comprehensive model of student learning that provides the analytical framework for the estimation of the variance of teacher quality.

3. THE IDENTIFICATION OF TEACHER EFFECTS

In this section we develop an estimator of the variance of teacher quality that avoids problems of student selection and administrator discretion that potentially have biased prior attempts. This estimator is based upon patterns of within-school differences in achievement gains and ignores variations in teacher quality across schools, because such variation cannot readily be disentangled from student differences and the contributions of other school factors. This strategy yields a lower bound estimator for the importance of teacher quality that relies upon minimal maintained assumptions about the underlying achievement process. Importantly, we do not focus solely on measurable characteristics of teachers or schools as is typically done in this literature but instead rely on student outcomes to assess the magnitude of total teacher effects, regardless of our ability to identify and measure any specific components. This semiparametric approach provides both an estimate of the role of teacher quality in the determination of academic achievement and information on the degree to which specific factors often used in determining compensation and hiring explain differences in teacher effectiveness.

3.1. *Basic Model of Student Achievement*

Academic achievement at any point is a cumulative function of current and prior family, community, and school experiences. A study of the entire process would require complete family, community, and school histories, and such data are rarely if ever available. Indeed, the precise specification of what to measure is poorly understood. In the absence of such information, analyses that study the contemporaneous relationship between the level of achievement and school inputs for a single grade are obviously susceptible to omitted variables biases from a number of sources.

An alternative approach focuses on the determinants of the *rate* of learning over specific time periods. The advantage of the growth formulation is that it eliminates a variety of confounding influences including the prior, and often unobserved, history of parental and school inputs. This formulation, frequently referred to as a value-added model, explicitly controls for variations in initial conditions when looking at how schools influence performance during, say, a given school year. While such a value-added framework by no means eliminates the potential for specification bias, the inclusion of initial achievement as a means to account for past inputs reduces dramatically the likelihood that omitted historical factors introduce significant bias.[4]

Equation (1) presents a conventional value-added model that describes the gain in student achievement (ΔA^{c}_{ijgs}) for individual i in cohort c with teacher j

[4]One restriction of this formulation is that the parameter estimates capture effects only for the specific period, ignoring any continuing impacts of inputs at an earlier age. See Krueger (1999) for a discussion of this issue. However, without detailed information and knowledge of the full

in grade g of school s:

$$(1) \qquad \Delta A^c_{ijgs} = A^c_{ijgs} - A^c_{ij'g-1s'} = X^c_{ig}\beta_X + T^c_{jgs}\beta_T + S^c_{gs}\beta_S + f_i + \varepsilon^c_{ijgs}.$$

This gain, measured as the difference between a student's test scores in grades g and $g-1$, depends on family background (X); teacher characteristics (T); school characteristics (S); inherent student abilities (f); and a random error (ε). Note that the term "inherent abilities" refers to the set of cognitive skills, motivation, and personality traits that affect the rate of achievement growth but that do not change during the school years being considered.[5] Each of the inputs can be thought of as a vector of underlying components.

Formulations similar to equation (1) have been estimated in a variety of circumstances in order to identify the causal link between a student outcome such as achievement or years of schooling on the one hand and a school characteristic such as class size on the other (see, e.g., Murnane (1975) or Summers and Wolfe (1977)). Much research has focused on the development of methods to eliminate any remaining biases, and we address this concern as well. However, a potentially much more important issue is the possibility that the measured teacher and school factors do not adequately capture important differences in the quality of education.

An alternative approach attempts to circumvent the problem of inadequate measures of quality through the estimation of classroom fixed effects on achievement gains (see, e.g., Hanushek (1971), Armor et al. (1976), Murnane and Phillips (1981)). These analyses of covariance capture all between-classroom differences in achievement gains controlling for any included regressors. The resulting classroom differences in average achievement gain have been interpreted as reflecting teacher quality, since the teacher is the most

cumulative achievement production process, it is virtually impossible to isolate any continuing effects of specific school factors.

The precise estimation approach found in the literature does vary. At times, initial achievement is added to the right-hand side of a regression equation, possibly with corrections for measurement error. At other times, simple differences or growth rates in scores are analyzed. The alternative formulations do place different restrictions on the form of the achievement process. See Hanushek (1979) for a discussion of value-added models. Subsequent analysis, relying on expected expansions of our database, will explore alternative specifications.

[5]The isolation of inherent student abilities does not rely on any presumption about their source (genetic, environmental, or an interaction of these). Any fixed differences that affect the rate of learning will be incorporated in this term. This formulation goes beyond typical discussions that concentrate just on how fixed ability, family, and motivational terms affect the level of achievement at a point in time. Here we explicitly allow for the possibility that *ceteris paribus* some children will acquire knowledge at different rates even after allowing for variations in initially observed achievement. Further, these differences do not have to be unidimensional.

obvious factor differing across classrooms. However, problems from test measurement errors and potential school and classroom selection effects may be even more serious for these types of models than in those that use observable measures, making the interpretation of these as direct estimates of the teacher component problematic.[6]

The central estimation problem results from the processes that match students with teachers, and schools. Not only do families choose neighborhoods and schools, but principals and other administrators assign students to classrooms. Because these decision makers utilize information on students, teachers and schools, information that is often not available to researchers or measured with error, the estimators are quite susceptible to biases from a number of sources. The following section develops an empirical model designed to avoid these problems and to identify the variations in the quality of instruction.

3.2. *An Extended Specification of Education Production*

Rather than attempting to define each variable in the education process, we begin by thinking in terms of the total systematic effect of students, families, teachers, and schools. In this, we depart from the parametric approach of equation (1) that involved measuring a small set of inputs in their natural units and move to a semiparametric approach with inputs measured in achievement, or output, units. Equation (2) describes a decomposition of education production during grade g into a set of fixed and time varying factors:

$$\Delta A^c_{ijgs} = \gamma_i + \theta_j + \delta_s + \upsilon^c_{ijgs}. \tag{2}$$

Test score gain in grade g is written as an additive function of student (γ), teacher (θ), and school (δ) fixed effects along with a random error (υ) that is a composite of time-varying components. The fixed student component captures the myriad family influences including parental education and permanent income that affect the rate of learning; the fixed school factor incorporates the effects of stable school characteristics including resources, peers, curriculum, etc. Finally, the teacher component captures the average quality of teacher j over time. Of course families, schools, and teachers all change from year to year, and such changes receive considerable attention in the analysis below.

Equation (2) is not intended to be a comprehensive model of the achievement determination process, and moreover we do not attempt to identify each of the separate components. Rather, it provides a framework for the specific models used to study the effects of teacher quality and school resource differences. We have not, for example, distinguished any role for school districts.

[6]Hanushek (1992) does provide suggestive evidence that teachers are the primary component by showing that classroom gains for individual teachers tend to be highly correlated across time (for different groups of students).

Many school policies—hiring, curriculum, school structure, etc.—emanate from school districts and will produce common elements in the teacher and school effects specified in equation (2). While the study of district effects is clearly important, particularly in a policy context, our focus on within-school achievement differences to avoid the difficulties associated with the endogeneity of school and district choice precludes identification of separate district effects.[7] Moreover, school fixed effects also capture any systematic differences across districts and communities, so there is no econometric reason to specify separate district or community components in this estimation. We do, however, address district related issues as they are relevant to the identification of teacher quality and school resource effects.[8]

3.3. *Estimator of the Variance of Teacher Quality*

In the semiparametric approach of equation (2), the variance of θ measures the variation in teacher quality in terms of student achievement gains. One could estimate this variance directly using between-classroom differences in average achievement gains. We do not adopt this approach for a number of reasons, not the least of which is the inability to match students to specific teachers. Yet even if students could be matched with teachers and the analysis considered only within-school variation in outcomes, both the intentional placement of students into classrooms on the basis of unobservables and the need to account for the contribution of measurement error to the between-classroom variation would introduce serious impediments to the identification of the variance of teacher quality.[9]

Consequently, we adopt a very different method that makes use of information on teacher turnover and grade average achievement gains to generate a lower bound estimate of the within-school variance in teacher quality. This approach avoids the need to identify and to estimate separately the test error

[7]The role of district environment and policies is a topic that we intend to pursue in the future. That analysis however, requires a different estimation strategy that, importantly, does not permit the precise identification of teacher influences that we pursue here.

[8]The model also imposes the assumption of additive separability in order to simplify the presentation. We explore the possibility that the magnitudes of school resource effects vary by student characteristics, allowing for the most commonly cited type of potential complementarity. In addition, we recognize that the matching of students and teachers likely affects the average rate of learning in a school, and the subsequent inclusion of school and school-by-grade fixed effects captures any differences that are maintained across our observation period.

[9]This discussion can be directly linked to prior estimation of classroom fixed effects, which develop classroom gains after conditioning on measurable characteristics of students or schools. See, for example, Hanushek (1971), Armor et al. (1976), and Murnane and Phillips (1981). In such cases, the interpretation of the individual and school components of equation (3) would relate directly to dimensions not captured by the included characteristics, and the test measurement errors would remain.

variance, and the aggregation to the grade level circumvents any problems resulting from classroom assignment.[10] The cost of this aggregation is the loss of all within grade variation in teacher quality and the inability to trace out the teacher quality distribution.

Equation (3) represents average achievement gain in grade g in school s for cohort c as an additive function of grade average student and teacher fixed effects, a school fixed effect, and the grade average error:

$$\overline{\Delta A_{gs}^{c}} = \overline{\gamma_{gs}^{c}} + \overline{\theta_{gs}^{c}} + \delta_s + \overline{\upsilon_{gs}^{c}}. \tag{3}$$

With two different cohorts of students (c and c'), we can compare average gains in the same grade:

$$\overline{\Delta A_{gs}^{c}} - \overline{\Delta A_{gs}^{c'}} = \left(\overline{\gamma_{gs}^{c}} - \overline{\gamma_{gs}^{c'}}\right) + \left(\overline{\theta_{gs}^{c}} - \overline{\theta_{gs}^{c'}}\right) + \left(\overline{\upsilon_{gs}^{c}} - \overline{\upsilon_{gs}^{c'}}\right). \tag{4}$$

Notice in equation (4) that all fixed school components from equation (3) drop out because they exert the same effect for both cohorts. These eliminated factors include fixed aspects of peers, school administration, technology, and infrastructure as they affect the *growth* in achievement, even if they are grade specific. They also include systematic (time invariant) sorting of teachers by school or district that comes from a district's salary or general attractiveness along with its standard teacher assignment practices. The difference in cohort average achievement gains is thus a function of the between-cohort differences in teacher quality (θ), in fixed student and family factors (γ), and an average error component that includes not only measurement errors but time varying individual, family, and school factors.

Though we do report estimates of the variance in teacher quality based on simple between-cohort achievement differences for a single grade, cohort average differences in (γ) contaminate estimates of the variance in teacher quality. Consequently, we concentrate on the difference between adjacent cohorts in the *pattern* of average gains in grades g and g'. In order to control fully for student fixed effects, we limit the sample to students who remain in the same school for grades $g-1$ and g:

$$\begin{aligned}\left(\overline{\Delta A_{gs}^{c}} - \overline{\Delta A_{g's}^{c}}\right) &- \left(\overline{\Delta A_{gs}^{c'}} - \overline{\Delta A_{g's}^{c'}}\right) \\ &= \left[\left(\overline{\theta_{gs}^{c}} - \overline{\theta_{g's}^{c}}\right) - \left(\overline{\theta_{gs}^{c'}} - \overline{\theta_{g's}^{c'}}\right)\right] + \left[\left(\overline{\upsilon_{gs}^{c}} - \overline{\upsilon_{g's}^{c}}\right) - \left(\overline{\upsilon_{gs}^{c'}} - \overline{\upsilon_{g's}^{c'}}\right)\right].\end{aligned} \tag{5}$$

[10]This estimator assumes that there are not strong complementarities between specific students and teachers, that is, that the effects of teachers is linear and separable as in equation (2). Yet as long as schools maintain similar assignment practices from year to year, as discussed below, even such complementarities will not contaminate the estimates. Additionally, changes in assignment practices will tend to bias estimates of the variance in teacher quality downward, reinforcing our interpretation of the estimator as a lower bound on teacher quality variance.

As equation (5) shows, taking the difference between average gains in grades g and g' eliminates all fixed student and family differences, leaving only cohort-to-cohort differences in the grade average difference in teacher quality and time varying student and school factors (contained in v) as determinants of the difference in the pattern of achievement gains.

Squaring both sides of equation (5) gives

$$
\begin{aligned}
(6)\qquad & \left[\left(\overline{\Delta A_{gs}^{c}} - \overline{\Delta A_{g's}^{c}}\right) - \left(\overline{\Delta A_{gs}^{c'}} - \overline{\Delta A_{g's}^{c'}}\right)\right]^2 \\
&= \overline{\theta_{gs}^{c}}^2 + \overline{\theta_{g's}^{c}}^2 + \overline{\theta_{gs}^{c'}}^2 + \overline{\theta_{g's}^{c'}}^2 - 2\left(\overline{\theta_{gs}^{c}}\,\overline{\theta_{gs}^{c'}} + \overline{\theta_{g's}^{c}}\,\overline{\theta_{g's}^{c'}}\right) \\
&\quad + 2\left[\left(\overline{\theta_{gs}^{c}}\,\overline{\theta_{g's}^{c'}} - \overline{\theta_{gs}^{c}}\,\overline{\theta_{g's}^{c}}\right) + \left(\overline{\theta_{g's}^{c}}\,\overline{\theta_{gs}^{c'}} - \overline{\theta_{gs}^{c'}}\,\overline{\theta_{g's}^{c'}}\right)\right] + e.
\end{aligned}
$$

The squared difference leads to a natural characterization of the observed achievement differences between cohorts as a series of terms that reflect variances and covariances of the separate teacher effects plus a component e that includes all random error and cross product terms between teacher and other grade specific effects.

We now impose three assumptions that formally characterize the notion that teachers are drawn from common distributions over the restricted time period of our cohort and grade observations: (i) The variance of grade average teacher quality is the same for all cohorts and grades; (ii) the covariance of grade average teacher quality for adjacent cohorts is the same for all grades; and (iii) the covariance of grade average teacher quality for grades g and g' for adjacent cohorts equals the covariance of grade average teacher quality for grades g and g' for each cohort. For ease of exposition, we also make the simplifying assumption that each school has one teacher per grade, but this is relaxed later.

Applying these assumptions and taking the expectation of equation (6) yields

$$
(7)\qquad E\left[\left(\overline{\Delta A_{gs}^{c}} - \overline{\Delta A_{g's}^{c}}\right) - \left(\overline{\Delta A_{gs}^{c'}} - \overline{\Delta A_{g's}^{c'}}\right)\right]^2 = 4\left(\sigma_{\theta_s}^2 - \sigma_{\theta_s^c \theta_s^{c'}}\right) + E(e_s),
$$

where $\sigma_{\theta_s}^2$ is the variance of teacher quality in school s and $\sigma_{\theta_s^c \theta_s^{c'}}$ is the covariance of teacher quality across cohorts in a school.

The key to the identification of the magnitude of the within-school variance of teacher quality comes from the first element on the right-hand side—the within-school variance of grade average teacher quality minus the within-school covariance of quality across cohorts. Consider first schools in which the two cohorts have the same teacher in each grade (i.e., the proportion of teachers who are different equals zero). As long as teachers perform equally well in both years, $\sigma_{\theta_s}^2 = \sigma_{\theta_s^c \theta_s^{c'}}$, and teacher quality contributes nothing to student performance *differences* across cohorts.

On the other hand, consider schools in which cohorts c and c' have different teachers in each grade (the proportion of teachers who are different

equals one).[11] In this case the within-school covariance of teacher quality equals zero. Importantly, this is not to say that schools hire randomly, for as we discuss below there can be little doubt that hiring practices and characteristics related to teacher job preferences differ substantially across schools. Rather, it says that the covariance across teachers in the deviation from the mean teacher quality in a school is zero.

Equation (7) provides the basis for estimation of the within-school variance of teacher quality. The left-hand side in most regressions is the squared divergence of the grade pattern in gains across cohorts, which we regress on the proportion of teachers who are different. Ignoring the possible confounding influences of other factors and maintaining the assumption that teacher quality remains unchanged in the absence of turnover, the coefficient on this proportion divided by four will provide a consistent estimate of the within-school variance in teacher quality.[12]

One empirical complication arises because most schools do not have a single teacher for each grade. Rather the number of teachers varies by school, and consequently the coefficient on the turnover variable would not have a straightforward interpretation. Because the achievement gains and the effects of teachers are averaged across the teachers in a grade, we actually have the variation of the mean in each school, and the relationship of turnover to the within-school variance will depend on the number of teachers. For example, in a sample of schools with three teachers per grade, the coefficient on proportion different would provide an estimate of four times one third (i.e., $4\sigma^2_{\theta_s}/3$) of the within-school variation in teacher quality. This also means that fifty percent turnover in schools with three teachers per grade would lead to the same expected squared cohort difference in grade average difference in gains as one hundred percent turnover in schools with six teachers per grade. In order to account for such differences in the number of teachers and place all schools on a common metric, the proportion differ-

[11]Note that such differences result from both teacher departures and grade changes. There is an extensive related literature on the determinants of teacher turnover, indicating that salary, working conditions, and alternative wage opportunities do affect the probability of exiting a school (cf. Dolton and van der Klaauw (1995, 1999), Murnane and Olsen (1989), Stinebrickner (2002), Hanushek, Kain, and Rivkin (2004b)). None, however, suggests that leavers are systematically more effective teachers than stayers, an issue to which we return below. Moreover, our analysis of within-school patterns of student performance implicitly controls for the overall determinants of turnover and focuses solely on the implications of turnover for performance. Regardless of any differences between leavers and stayers, the within-school covariance of grade average quality equals zero in 100 percent turnover schools as long as any changes in hiring procedures are not systematically related with the quality of leavers.

[12]Note that we use teacher turnover as a method of identifying the variance in teacher quality. Implicitly, we assume teacher turnover does not directly affect student achievement gains except for the possibility of systematic quality differences by teacher experience. We test this assumption within the general production function estimation (below) and cannot reject it.

ent must be divided by the number of teachers per grade, and the coefficient on this variable provides an estimate of the within-school variance in teacher quality.

Our empirical strategy focuses on the estimation of a lower bound on the variation of teacher quality, and in that regard a variety of factors that suggest downward bias in our turnover estimator are not problematic. First is the almost certain violation of the assumption that the variance and covariance terms are equal in schools without turnover. Even in the absence of teacher turnover, there is almost certainly some difference in teacher quality from year-to-year due to changes in pedagogy, personal problems, learning (particularly for beginning teachers), etc., reducing the expected coefficient on the turnover variable below the true within-school variance.

Measurement error in the teacher turnover variable would tend to exacerbate any such downward bias. The administrative data have missing information on key variables, and it is not always clear who teaches which subjects. Consequently, there is some error introduced into the calculations of both the percentage of teachers who differ from cohort to cohort and in the number of teachers per grade, and the ratio of the two may thus contain a nontrivial amount of noise.

More worrisome for our approach, however, is that there are also two potentially important sources of upward bias. First is the standard problem of omitted variables. Teacher turnover may be precipitated or accompanied by other changes such as a new principal or superintendent or district induced curriculum changes (Ingersoll (2001)). If, for example, administrator turnover also leads to teacher turnover, any direct effects of new administrators on achievement growth would introduce an upward bias if they were not accounted for. In the empirical work below, we take a number of steps to control for potentially confounding time-varying factors including controls for the numbers of principal and superintendent changes over the observation period. We also perform various sensitivity analyses directed at these issues.

Second is the possibility that teachers who exit are not drawn randomly from the teacher quality distribution. If attrition and quality are systematically related, the average teacher quality in high turnover years will tend to differ systematically from the average quality of new hires. Consider the possibility that high quality teachers are more likely to exit. In this case, schools that obtain a particularly good draw of teachers in one year will tend to experience both a greater turnover following the year and a larger average difference in achievement gains than would be experienced with random attrition. This situation would lead to an upward bias in our estimator, as would the opposite case where low quality teachers are more likely to exit. Even if attrition and quality are uncorrelated, if teachers in the tails of the distribution are more likely

to exit, higher turnover schools will tend to have higher cohort differences in achievement gains, again biasing our estimator upward.[13]

Appendix A demonstrates that a major departure from random exiting in the form of higher probabilities in either or both tails of the distribution can introduce substantial upward bias. In the absence of student/teacher matches, we have little information on the actual distribution of departures. Moreover, the literature on teacher turnover is not very informative on the quality distribution of any school attrition.[14] A general presumption, particularly in more policy-related analyses, is that union restrictions, the single salary schedule for teachers, and the lack of performance incentives related to student achievement mute any relationship between teacher quality and attrition, but this is clearly speculative.[15] Fortunately, we do have student/teacher matches for a single district, and we use that information to provide empirical evidence on the likely magnitude and direction of any nonrandom turnover induced bias.

Finally, this framework relies on just the variation in teacher quality that is found within schools and ignores all variation in teacher quality across schools. If all schools were to hire randomly from a common pool, the between-school variance would equal zero, but this is almost certainly not the case. Rather schools able to offer higher salaries or better working conditions choose among a larger pool of applicants and likely enjoy higher average teacher quality, though the difficulty predicting productivity on the basis of education credentials and interviews almost certainly allows for substantial within-school heterogeneity.[16] In the extreme, if schools were perfectly arrayed in their hiring, all variations in quality would be between schools. In any event, the between-school differences would have to be added to the estimates reported below to obtain an estimate of the total variation in the quality of instruction.

[13]Note that heavy attrition in just one tail also implies drift in the average quality of teachers, which would inappropriately add to our estimate of the within-school variance (and which we explicitly assume is not the case).

[14]Much of the turnover literature (footnote 11) relates to opportunity costs by specialties (e.g., math and science), but these studies are more relevant for secondary schools and do not directly address issues of quality. Another approach investigates attrition by the teacher's own test score (see Murnane et al. (1991)) and finds some relationship suggesting that higher scoring teachers are more likely to leave, but neither this relationship nor the relationship between teacher test scores and student achievement is very strong. The one direct study relating attrition to classroom performance finds that principal evaluations early in the teaching career are positively correlated with continued teaching. At the same time, while teacher value-added based on student achievement is also positively related to retention of teachers, the estimates are statistically insignificant (Murnane (1984)), perhaps because of the small samples.

[15]For example, The Teaching Commission (2004, p. 46) notes: "once teachers have passed a probationary period, it is notoriously difficult to dismiss those whose performance is inadequate. In 2002, for instance, only 132 of 78,000 teachers in New York City's massive school system were removed for poor performance." However, no analyses of decisions before tenure or of more informal actions are available.

[16]Hanushek, Kain, and Rivkin (2004b) find that teachers who switch schools tend to move to schools with higher achieving, higher income, and lower proportion minority student bodies.

4. THE TEXAS DATABASE

The data used in this paper come from the UTD Texas Schools Project, conceived of and directed by John Kain. Data are compiled for all public school students from administrative records in Texas, allowing us to use the universe of students in the analyses. We use data for three cohorts: 3rd through 7th grade test scores for one cohort (4th graders in 1995) and 4th through 7th grade test scores for the other two (4th graders in 1993 and 1994).[17] For each cohort there are more than 200,000 students in over 3,000 public elementary and middle schools. (For details on the database, see Appendix B and Table B1; currently available data along with variable definitions and estimation programs are found in Rivkin, Hanushek, and Kain (2005).) In comparison to studies that use only a small sample of students from each school, these data permit much more precise estimates of school average test scores and test score gains.

The administrative data contain a limited number of student and family characteristics including race, ethnicity, gender, and eligibility for a free or reduced price lunch. Students who switch public schools anywhere within the state of Texas can be followed just as those who remain in the same school or district. Although explicit background measures are relatively limited, the panel feature can be exploited as described previously to account implicitly for time invariant individual and school effects on achievement.

Beginning in 1993, the Texas Assessment of Academic Skills (TAAS) was administered each spring to eligible students enrolled in grades 3 through 8.[18] These tests are designed to evaluate student mastery of the grade-specific subject matter that is prescribed for students in the state.[19] We focus on test results for mathematics and reading, derived from tests of approximately fifty questions. Because the number of questions and average percent correct varies across time and grades, we transform all test results into standardized scores with a mean of zero and variance equal to one, though the empirical findings

[17]Note that, while we have 3rd grade test information, our analysis begins at 4th grade because of the focus on achievement gains.

[18]Many special education and Limited English Proficiency (LEP) students are exempted from the tests, as are other students for whom the test would not be educationally appropriate. In each year roughly fifteen percent of students do not take the tests, either because of an exemption or because of repeated absences on testing days. This rate of missing tests appears comparable to those for other high quality testing programs such as the National Assessment of Educational Progress.

[19]The TAAS tests are generally referred to as criterion referenced tests, because they refer directly to pre-established curriculum or learning standards. The common alternative is norm referenced tests that cover general subject matter appropriate for the subject and grade but that are not as closely linked to the specific state teaching standards. In principle, all students could achieve the maximum score on a criterion referenced test with no variation, while norm referenced tests focus on obtaining information about the distribution of different skills across the tested population. In practice, scores on commonly available criterion referenced and norm referenced tests are highly correlated across students.

are robust to a number of transformations including the raw percentage correct. The bottom one percent of test scores (all less than or equal to expected scores from random guesses) are trimmed from the sample in order to reduce measurement error. Participants in bilingual or special education programs are also excluded from the samples used in estimating teacher quality and resource effects because of the difficulty in measuring teacher and school characteristics for these students.[20]

Student data are merged with grade average information on teachers by subject. Because student and teacher data come from different reporting systems that are not directly linked, matching students with their specific teachers is not possible. Teacher personnel data provide information on experience, highest degree earned, and the class size, subject, grade, and population served for each class taught. This information is used to construct subject and grade average characteristics for teachers in regular classrooms. In the early grades teachers tend to teach all subjects, while in junior high most specialize. We consider those who self identify as general teachers as teachers of both mathematics and reading.

5. LOWER BOUND ESTIMATES OF THE VARIANCE OF TEACHER QUALITY

The estimation of the within-school variance in teacher quality relies on the notion that teacher turnover increases the variance in student outcomes across grades and cohorts in a school. Although we refine the estimation below, the pattern can be seen directly by observing the higher correlations in student achievement across cohorts for schools with lower teacher turnover (fewer than twenty five percent of teachers are different) than schools with high turnover (fewer than twenty five percent of teachers are the same). The correlations are 0.40 in math and 0.26 in reading respectively for the low turnover schools and 0.22 in math and 0.14 in reading for the high turnover schools. Of course other factors correlated with teacher turnover could also produce this pattern, and it is necessary to turn to our more structured model in order to identify the importance of teacher quality in the determination of achievement gains. Note that on average roughly one third of teachers are new to a grade and subject in any year. This is roughly double the rate of school leaving, meaning that incumbent teachers tend to change grades or subjects every five years or so.

[20]For an explicit analysis of the achievement of special education students, see Hanushek, Kain, and Rivkin (2002). Kain and O'Brien (1998) provide additional analysis of special education students along with information on the performance of limited English proficiency (LEP) students. These students are included in the calculations of class sizes for the analysis below when they receive instruction in regular classrooms.

5.1. *Basic Estimates*

Table III reports basic estimates from the regression of the squared between-cohort difference in gains on the proportion of teachers who are different and other covariates. The sample includes only students who remain in the same school for two successive grades, either 5th and 6th or 6th and 7th, and only grades that have at least five students with valid test scores and nonmissing data on teacher turnover.[21] Just grades 5 and 6 are used for the small number of schools with all three grades.[22] The final sample has 3,076 schools in the mathematics specifications and 3,086 in the reading specifications.

The three left hand columns in Table III report results from the three specifications for mathematics and reading in order to isolate the sensitivity of the estimates to the different fixed components of achievement growth. The first regresses the squared difference in 5th (or 7th) grade gains between cohorts on 5th (or 7th) grade teacher turnover; the second and third regress the squared difference in the difference of 5th (or 7th) and 6th grade gains between cohorts on the turnover of 5th (or 7th) and 6th grade teachers combined. As described previously, using the difference in gains between the two grades controls for both student and school fixed effects in gains. Finally, the third specification adds an additional school fixed effect directly into the regression, identifying the variance in teacher quality on the basis of the difference in turnover rates between the first and second cohorts and the second and third cohorts. This last estimation, which captures school specific variations in the grade pattern of performance, directly controls for systematic school and grade specific unobservables that may be correlated with turnover. All three specifications also include a dummy variable identifying the precise cohort comparison, the inverse of enrollment (because the variance of measurement error in student performance is inversely proportional to enrollment), the use of 7th grade information, and the numbers of new principals and superintendents. The measures of new school and district leadership capture time varying policy factors that could simultaneously affect teacher turnover and student achievement.

The results show that differences in mathematics and reading achievement gains among cohorts are strongly related to teacher turnover. All coefficients

[21] An additional observation in the reading sample was also excluded, because the grade average gain was more than six standard deviations from the mean (higher than any other school). It turned out to be a single teacher whose students' average gain in the previous year was quite close to the mean and who did not teach in the subsequent year. In addition, the average gain in the subsequent grade was roughly four standard deviations below the mean, far different than the positive gain reported for the prior cohort taught by the same teacher. We believe there is overwhelming evidence of either cheating or miscoding. The exclusion of this observation did not have a large impact on the estimates except in the full fixed effect model.

[22] The majority of students move from elementary to middle school sometime between grades 5 and 7. Roughly fifteen percent of schools with at least two of the three grades in this range have all three.

TABLE III

EFFECT OF TEACHER TURNOVER ON THE DIVERGENCE OF MATHEMATICS AND READING TEST SCORE GAINS BETWEEN COHORTS (STANDARD ERRORS IN PARENTHESES)

	No Fixed Effects[a]	Individual and School Fixed Effects[b]	Individual and School-by-Grade Fixed Effects[c]	Individual and School Fixed Effects[b]	Individual and School-by-Grade Fixed Effects[c]
1. Mathematics					
Proportion of teachers who are different/number of teachers	0.080 (0.017)	0.090 (0.015)	0.050 (0.021)	0.080 (0.016)	0.045 (0.021)
Absolute change in proportion of teachers with no experience				0.033 (0.016)	0.027 (0.023)
2. Reading					
Proportion of teachers who are different/number of teachers	0.067 (0.013)	0.082 (0.014)	0.036 (0.018)	0.078 (0.015)	0.029 (0.018)
Absolute change in proportion of teachers with no experience				0.015 (0.015)	0.041 (0.020)

Notes: All equations include the inverse of the number of students, numbers of new principals and superintendents in the school during adjacent years, a grade 7 dummy variable, and a cohort dummy variable. The sample includes all students who remain in the same school for grades 5 and 6 (or 6 and 7). Sample size is 3,076 for the mathematics and 3,086 for the reading specifications.

Equations have the same structure for mathematics and for reading. (The analyses of gain patterns between grades 6 and 7 take the same form as those for grades 5 and 6 that are shown.) For Φ = proportion different math (or reading) teachers/#teachers and adjacent cohorts (c and c'), the specifications take the following forms:

(a) $(\bar{A}_5^c - \bar{A}_5^c)^2 = \beta_\theta \Phi_{6s}^{c,c'} + \beta_X X_{6s}^{c,c'} + e_{6s}^{c,c'}$,

(b) $[(\bar{A}_6^c - \bar{A}_5^c) - (\bar{A}_6^{c'} - \bar{A}_5^{c'})]^2 = \beta_\theta \Phi_{5 \text{ and } 6,s}^{c,c'} + \beta_X X_{5 \text{ and } 6,s}^{c,c'} + e_{5 \text{ and } 6,s}^{c,c'}$,

(c) $[(\bar{A}_6^c - \bar{A}_5^c) - (\bar{A}_6^{c'} - \bar{A}_5^{c'})]^2 = \beta_\theta \Phi_{5 \text{ and } 6,s}^{c,c'} + \delta_s + \beta_X X_{5 \text{ and } 6,s}^{c,c'} + e_{5 \text{ and } 6,s}^{c,c'}$,

where δ_s is a fixed effect for school s.

are positive and significant at the five percent level, and except for the school-by-grade fixed effect specifications, all t-statistics exceed 4.5 in absolute value. The declines in coefficient magnitudes for the full fixed effect specifications are consistent with measurement error induced attenuation bias, but they may also reflect the presence of omitted variables bias in the other specifications. In order to avoid as much as possible the introduction of any upward biases, we concentrate here on the full fixed effect coefficients of 0.050 and 0.036. These imply lower bound estimates of the within school variance of teacher quality (measured in units of student achievement) equal to 0.0125 (0.050/4) and 0.009 (0.036/4) for mathematics and reading respectively. This means that a one standard deviation increase in average teacher quality for a grade raises average student achievement in the grade by at least 0.11 standard deviations of the total test score distribution in mathematics and 0.095 standard deviations in reading.

These estimates suggest the existence of substantial within school variation in teacher quality, but they combine average differences across the experience distribution with skill differences not related to experience. As we demonstrate in the direct estimation of educational production functions below, the learning curve appears to be quite steep in the first year or two of teaching before flattening out. Because many of the teachers new to a grade are in their first year, the share of the variance due to differences between beginning and experienced teachers might be quite sizeable. Fortunately, we can identify the effects of beginning teachers by including the absolute change in the share of teachers in their first year as an additional variable.[23]

The final two columns of Table III present estimates from the two fixed effect specifications that include the absolute change in the share of beginning teachers. These estimates suggest that quality differences between new and experienced teachers account for only ten percent of the teacher quality variance in mathematics and somewhere between five and twenty percent of the variance in reading. The addition of the change in the share of teachers with one year of experience (not shown) has virtually no effect on the estimates.

5.2. *Specification Checks*

The consistency of the estimator relies on the assumption that the turnover variable is unrelated to the error. One important threat to the estimation strategy is the possibility that unobserved changes over time in schools may be correlated with teacher turnover. A comprehensive control for other time varying factors in the schools comes from looking at turnover of teachers not involved in the specific subject. Specifically, by looking at schools that use separate teachers for mathematics and English, we can include English teacher turnover as a control variable in the modeling of math performance and mathematics teacher turnover in the modeling of reading achievement.[24]

Table IV reports the results for fixed effect specifications that include turnover in the untested subject. These estimates are generated from the smaller subsample of schools with subject specialists (defined as schools that have no teachers in either of the two sampled grades who teach both math and English), which is roughly thirty percent of the full sample. The results for mathematics remain highly significant though somewhat smaller in the first two specifications and are significant only at the ten percent level in the full fixed effects model, which is not that surprising given the substantial reduction in

[23] Because we are looking at variance in outcomes across cohorts, any significant change either up or down in the proportion of teachers in their initial year of experience has a similar impact, thus making the absolute value appropriate.

[24] Because teacher turnover in the untested subject is used to identify any concomitant disruption in the school, the number of teachers in that subject will not directly affect the variance in student performance. Therefore this turnover variable is not divided by the number of teachers in the untested subject.

TABLE IV

EFFECT OF TEACHER TURNOVER ON THE DIVERGENCE OF MATHEMATICS AND READING TEST SCORE GAINS BETWEEN COHORTS, CONTROLLING FOR TEACHER TURNOVER IN OTHER SUBJECTS[a] (STANDARD ERRORS IN PARENTHESES)

	Individual and School Fixed Effects[b]			Individual and School-by-Grade Fixed Effects[b]		
1. Mathematics						
Proportion different math teachers/number of teachers	0.059 (0.015)	0.058 (0.015)	0.069 (0.016)	0.034 (0.021)	0.034 (0.021)	0.035 (0.021)
Absolute change in proportion math teachers with no experience			−0.029 (0.013)			−0.005 (0.020)
Proportion of same English teachers		−0.006 (0.010)	−0.008 (0.010)		0.002 (0.014)	0.002 (0.014)
2. Reading						
Proportion different English teachers/number of teachers	0.027 (0.016)	0.024 (0.016)	0.010 (0.016)	0.001 (0.021)	−0.001 (0.021)	−0.005 (0.022)
Absolute change in proportion English teachers with no experience			0.042 (0.015)			0.013 (0.021)
Proportion of same mathematics teachers		−0.017 (0.011)	−0.016 (0.011)		−0.020 (0.013)	−0.020 (0.013)

[a]The sample includes all students who remain in the same school for grades 5 and 6 (or 6 and 7) in schools with no teacher offering both English and math instruction. All equations include the inverse of the number of students, numbers of new principals and superintendents in the school during adjacent years, a grade 7 dummy variable, and a cohort dummy variable. The sample size is 855.

[b]Table III notes describe the estimation specifications.

sample size. In contrast, the English teacher turnover coefficients in the reading test score regressions become quite small and insignificant in all specifications, raising concern that confounding factors in this estimation method could be driving the results. In this sample, the impact of inexperienced teachers is very imprecisely estimated. Importantly, comparisons across specifications for a common sample reveal that the inclusion of turnover information for the untested subject has virtually no effect on the other turnover estimate in either fixed effect specification.

The question remains as to why the estimates in Table IV are uniformly smaller than those reported in Table III. An important difference between the samples for the respective tables is the balance between 5th and 7th grade classrooms. It is almost always the case that junior high schools use subject specific teachers, while elementary schools use a single teacher for most subjects. Consequently the vast majority of schools with subject specific teachers include grades 6 and 7, while the majority of all schools in the sample include grades 5 and 6. Systematic differences by grade in the effects of teachers on test scores could therefore account for the observed pattern of results.

Table V reports estimates that allow the effect of turnover to vary by grade combination based on the full sample used in Table III. The coefficients suggest that the variance in teacher quality declines in mathematics as students progress through school, though the interaction term becomes insignificant in the full fixed effect model. On the other hand, it appears that within school differences in teacher quality are quite substantial in reading in elementary school but explain little or none of the variation in outcomes in junior high. In both subjects the pattern of estimates in Table V explain the differences between Tables III and IV. Interestingly, this pattern of diminishing effects will repeat itself in the production function estimates below, suggesting either that school

TABLE V

GRADE DIFFERENCES IN THE EFFECTS OF TEACHER TURNOVER ON THE DIVERGENCE OF MATHEMATICS AND READING TEST SCORE GAINS BETWEEN COHORTS (STANDARD ERRORS IN PARENTHESES)[a]

	Individual and School Fixed Effects	Individual and School-by-Grade Fixed Effects	Individual and School Fixed Effects	Individual and School-by-Grade Fixed Effects
1. Mathematics				
Proportion of teachers who are different/number of teachers	0.113 (0.018)	0.063 (0.026)	0.096 (0.019)	0.036 (0.027)
6th & 7th grades interaction with proportion different[b]	−0.075 (0.031)	−0.036 (0.044)	−0.068 (0.032)	−0.024 (0.047)
Absolute change in proportion of teachers with no experience			0.026 (0.022)	0.052 (0.032)
6th & 7th grades interaction with absolute change[b]			0.018 (0.033)	−0.035 (0.048)
2. Reading				
Proportion of teachers who are different/number of teachers	0.115 (0.017)	0.066 (0.022)	0.114 (0.018)	0.059 (0.023)
6th & 7th grades interaction with proportion different[b]	−0.092 (0.028)	−0.081 (0.037)	−0.101 (0.029)	−0.083 (0.038)
Absolute change in proportion of teachers with no experience			0.004 (0.020)	0.048 (0.027)
6th & 7th grades interaction with absolute change[b]			0.030 (0.031)	−0.011 (0.039)

[a]Table III notes describe the sample and estimation specifications.

[b]Interaction between an indicator for the grade 6 and 7 observations and specified variable.

and teacher quality differences have much smaller effects on achievement in junior high or that the test results do a poor job of capturing differences in school quality in those grades.

There remains one other potential source of bias that must be addressed. Although controls for any concomitant changes to teacher turnover address the problem of omitted variables, they do not resolve the potential problem of nonrandom teacher attrition described above. As noted previously, the estimation relies upon the assumption that turnover is uncorrelated with quality and is not drawn heavily from either of the tails of the quality distribution. Since our estimator is identified by the assumption of random departures, we cannot readily test this assumption within our model and data.

Fortunately, for one large Texas school district we have developed some additional data that link student test score gains with individual teachers.[25] Although we cannot account for unobservable selection into classes, sampling error, and the other factors that we explicitly worry about in this paper, we can use these data to compute a within-school measure of quality: average student achievement gains for each teacher minus the average for all teachers in the same school that year. We can then calculate attrition probabilities based on this quality measure and use these probabilities to estimate the impact of any nonrandom attrition on our estimator of the variance of teacher quality.

Table VI describes the distribution of teachers placed into twenty quality categories along with the probabilities of exit for each group. We create these categories by dividing the range of teacher average gains relative to the school average into twenty intervals of equal length. (Because of concerns about outliers, we drop the top and bottom one percent of gains, but the results are invariant to this sampling procedure as we show below.) Within each category we use the mean gain as the index of quality. Since the division into twenty categories is arbitrary, we examine the sensitivity of the results to changes in the number of intervals.

With random departures there would be no systematic differences in the probability of exiting. This does not appear to be the case in Table VI, as attrition clearly declines with quality, probably in part due to the fact that first year teachers have the highest attrition. On the other hand, attrition does not appear to be concentrated in the tails of the distribution, the key element described in Appendix A. (Note that there are very few teachers in the lowest quality category that is an outlier in the exit rate at 42.9 percent.)

We now use the method developed in the simulations in Appendix A to estimate the bias introduced by deviations from random departure of the type observed in Table VI. Table C1 shows that the nonrandom attrition leads to a very slight increase (less than one percent) in the estimated standard deviation of teacher quality. This result also holds if the number of quality intervals is doubled or tripled or if observations in the tails of the distribution are retained

[25]These data are described in Hanushek et al. (2005).

TABLE VI

TEACHER EXIT RATES BY QUALITY OF INSTRUCTION RELATIVE TO OTHERS IN THE SCHOOL FOR TEACHERS IN A LARGE TEXAS DISTRICT

Quality Index	Frequency (Percent)	Exit Rate (Percent)
−1.56	0.17	42.9
−1.41	0.20	11.8
−1.27	0.45	23.7
−1.11	0.56	23.4
−0.94	1.17	30.6
−0.79	1.73	26.2
−0.63	2.86	22.2
−0.48	5.08	22.6
−0.32	9.58	21.3
−0.16	15.29	20.6
−0.01	21.35	20.2
0.14	16.65	17.65
0.29	10.58	18.51
0.45	6.51	18.35
0.60	3.55	12.79
0.76	2.07	17.34
0.92	0.96	25.00
1.07	0.62	13.46
1.22	0.43	13.89
1.38	0.19	0.00

Notes: The sample includes all teachers in grades 4–8 in one large Texas district. The measure of quality is the difference between average student gain in mathematics for a teacher and the average gain for all other teachers in the school. These relative gains are divided into twenty equal intervals, and the index for each interval is the interval mean. Frequency is the percentage of all teachers in the city in the category, and exit rate is the percentage of teachers who leave the school at the end of the year.

in the sample. Therefore, even if attrition is not random for the sample as a whole, as long as it is not far more concentrated in the tails than is observed for this single large district, it is extremely unlikely that it would introduce much if any upward bias.[26]

A final robustness check examines only schools with a single teacher per grade. This quite select sample generates large, positive, and statistically significant estimates in both mathematics and reading for the first two specifications (see Table C2). Not surprisingly given the extremely small sample sizes, the estimates for the full fixed effect specification remain positive but are quite imprecise.

[26]Note that the estimates of within school variation in quality based on individual teachers are three times as large as our lower bound estimates in Table III. Of course, these estimates do not deal with the selection effects that are the heart of the estimation here. They also include potentially important measurement error.

Importantly, the true magnitudes of the variances in mathematics and reading teacher quality are likely to be larger than the estimates presented here. First, the identifying assumptions are likely to be violated in ways that bias downward the extent of actual teacher quality differences within schools. Second, the measures of teacher turnover and number of teachers likely contain some error, and the ratio of the two may in fact have substantial measurement error that would likely attenuate the coefficients. For example, the exclusion of schools with large changes in the number of teachers in a grade from year to year, an indicator of problematic data, tends to increase coefficient magnitudes and the precision of virtually all estimates. Finally, we focus on just one component of the variance in teacher quality, the within-school variance. All between-school variation in teacher quality is ignored—not because of a belief it is small, but rather because it cannot be readily separated from other factors. Thus, there can be little doubt that teacher quality is an important determinant of reading and mathematics achievement in elementary school and mathematics achievement in junior high school.

6. EDUCATION PRODUCTION FUNCTION ESTIMATES

The frequently employed implicit assumption that schools are homogenous institutions is clearly contradicted by the finding of substantial within-school heterogeneity in teacher quality. These results also contrast sharply with the much smaller estimated differences in teacher and school quality that comes from studies investigating the impacts of specific school or teacher characteristics. Nevertheless, because teacher salaries are closely linked with experience and formal education and because class size reductions have been a widely discussed and often used policy tool, a better understanding of the effects of these specific factors remains important. From a policy viewpoint, a comparison of the costs and benefits of smaller classes or more educated and experienced teachers with those of improved general teacher quality would be particularly informative.

The results from the existing large body of literature on the effects of school resources on a variety of outcomes remain highly variable, in large part, we believe, because of difficulty of controlling for other relevant achievement inputs due to both conceptual and data limitations.[27] The main concern is that either explicit resource allocation rules—such as the provision of compensatory funds for poor achievers—or simple omitted variables problems could mask

[27]For summaries of the education production function literature, see Hanushek (1986, 2003), Levačić and Vignoles (2002), and Woessmann (2004). This work has been quite varied and controversial (Burtless (1996)). While concentrated on analyses of test score performance, continuing attention has also turned to longer run impacts on labor market outcomes (see, e.g., Card and Krueger (1992), Betts (1995), Heckman, Layne-Farrar, and Todd (1996), Dearden, Ferri, and Meghir (2002), and Dustmann, Rajah, and van Soest (2003)).

or distort true causal impacts. A set of more recent studies focuses specifically on identifying factors leading to exogenous variation in class size in order to uncover causal impacts.[28] Unfortunately, identification of truly exogenous determinants of class size, or resource allocations more generally, is sufficiently rare that other compromises in the data and modeling are frequently required. These jeopardize the ability to obtain consistent estimates of resource effects and may limit the generalizability of any findings.

As described in Section 3, our framework eliminates directly the most troubling potential endogeneity problems that are the focus of the alternative instrumental variables approaches. The large samples also permit detection of small effects that may differ by grade or student demographic characteristics, allowing us to distinguish between low power of tests and the true lack of a relationship.

6.1. *Empirical Specification of Resource Models*

Equation (8) describes the value-added empirical model that forms the basis of our examination of school resource effects on achievement. This is a modified version of equation (2) that adds a vector of school resource characteristics (SCH) measured at the grade level and a set of observable, time varying family characteristics (X):

$$\Delta A^c_{ijgs} = SCH^c_{gs}\lambda + X^c_{ig}\beta + \underbrace{\gamma_i + \delta_{sy} + \omega_{gs} + \upsilon^c_{ijgs}}_{\text{composite error}}. \tag{8}$$

The family characteristics include indicator variables for students who switch schools and students who are eligible to receive a free or reduced price lunch. Teacher and school characteristics are computed separately for each grade and subject, and they include the average class size in regular classrooms,[29] the proportion of teachers with a master's degree, and the proportion of teachers who

[28]A variety of different approaches have been applied to sort out the causal influence of school resources including instrumental variables approaches relying upon various circumstances of the schooling institutions (e.g., Angrist and Lavy (1999), Feinstein and Symons (1999), Hoxby (2000), Woessmann and West (forthcoming), Dobbelsteen, Levin, and Oosterbeek (2002), Robertson and Symons (2003), and Bonesrønning (2004)) and direct consideration of potential pre-treatment selection factors (e.g., Dearden, Ferri, and Meghir (2002)).

[29]As Boozer and Rouse (1995) and others have pointed out, it is important to separate regular and special education students, because class size and possibly other characteristics differ dramatically by population served and because special education students are much less likely to take tests. If the proportion of students in special education classes or the gap between regular classroom and special education class size differs across schools, estimates of the effect of class size based on the entire school average will be biased. Our measure of class size is the average class size for regular classrooms in specific grades and subjects. Both special purpose classes and student achievement for special education and Limited English Proficiency (LEP) students are eliminated from this estimation. At the same time, special education students in regular classroom instruction are included in the calculation of class size because they will affect the resources

fall into four experience categories: zero years, one year, two years, and three or four years (with the omitted category being five years and above).[30] The composite error terms should be reinterpreted as the unobserved components of students and schools. Note that we have added two additional error terms: school-by-year fixed effects (δ_{sy}) and school-by-grade fixed effects (ω_{gs}). These absorb the school fixed effects previously considered.

Unlike most educational studies, we concentrate specifically on the actual class sizes reported by regular classroom teachers rather than the more common pupil-teacher ratios for a school. Further, considerable attention was given to the elimination of measurement error in the school variables. We have access to longitudinal information on key data and can therefore adjust reports for inconsistencies that occur over time. Data Appendix B describes in detail the construction of the school characteristics and sample selection criteria.

Virtually all prior analyses of school resource effects have estimated specifications similar to equation (8) in either level or growth form, but none has been able to account for all of the fixed components of the composite error term. The elimination of these factors in the estimation of equation (8) addresses virtually all of the concerns typically raised about estimation of educational production functions. For example, arguments about simultaneity arising from compensatory resource allocations based on student performance are directly eliminated, since the level and expected rate of gain of achievement for each student are explicitly dealt with through the investigation of ΔA and the estimation of the individual γ_i's. The removal of school fixed effects would also control for time invariant school characteristics that might be related to the included teacher and school characteristics.

Though the removal of simple school fixed effects (δ_s) would eliminate the confounding influences of fixed school factors including stable curriculum, neighborhood factors, peer characteristics, school and district leadership, and school organization, changes over time in other school factors may be correlated to changes in the included teacher and school characteristics. Consider the possibility that other events in a school—leadership changes, curricular developments, student perceptions and flows, or the like—influence achievement directly and are correlated with changes in school and teacher characteristics. Importantly, the availability of a number of cohorts permits the inclusion of school-by-year fixed effects (δ_{sy}) rather than simple school fixed effects in some

allocated to regular instruction students in those classrooms. Separate analysis of special education is found in Hanushek, Kain, and Rivkin (2002).

[30]Including the percentages of teachers with five to nine and twenty or more years of experience as separate categories did not change any of the results, and the hypotheses that teachers with five to nine or twenty or more years of experience had a different impact from those with ten or more years of experience was rarely rejected at any conventional significance level. The class size and teacher education estimates also remained unchanged if average experience was used in place of the experience categories.

specifications in order to account for any such systematic year-to-year changes in school factors. Any pattern of events or policies common to the neighborhood and school will be eliminated, and the estimates are identified solely by within-school-by-year differences across grades.[31]

We believe an extremely strong case can be made that the remaining differences in class size and other teacher characteristics emanate from two uncontaminated sources: random differences between cohorts in the number of students who transfer in or out of the school as students age (i.e., changes in enrollment);[32] and school or district induced changes in class size policies that are unlikely to be systematically related to the time varying error components of individual students, controlling for student and school-by-year fixed effects in achievement gains.[33]

This approach to estimation goes well beyond what has been possible even with the specialized effects of institutional structure that have entered into past instrumental variables estimation. A concern, however, is that the signal to noise ratio falls with the removal of the multiple fixed effects, thus making it difficult to estimate the remaining elements of the specification. We consider this possibility below.

6.2. *Impact of Teacher and School Characteristics*

Table VII reports the full range of estimates obtained from value-added models that progressively contain no fixed effects; student and school fixed effects; student and school-by-year fixed effects; and, finally, student, school-by-year, and school-by-grade fixed effects.[34] Based on preliminary findings, class size effects are further allowed to differ by grade. Robust standard errors that account for the correlation of unobservables within a school are reported for all coefficients.[35] Table B1 presents descriptive statistics for the school characteristics and achievement gain.

[31] Less substantively, we also allow for changes in the tests over time through inclusion of a fixed effect for year for each subject-grade test (τ_{gy}).

[32] Note that the estimation explicitly controls for the effects of moving on the moving students' achievement growth; see Hanushek, Kain, and Rivkin (2004a).

[33] The availability of multiple cohorts also permits the inclusion of school-by-grade fixed effects, though at a cost of losing the ability to identify variable effects in the single 4th grade cohort. This may be important if, as suggested to us by Caroline Hoxby, school average achievement and class size change in a systematic way as students progress through school. However, the lack of systematic differences in class size by student demographic composition in any grade suggests that such problems are very minor if they exist at all. In the most complete model, coefficients are identified by school-by-grade-by-year differences in characteristics and achievement gains.

[34] Related to the work in the prior section, we also included (not shown) the level of teacher turnover in each year but found that it never had a systematic influence on student achievement. Stable differences in teacher turnover for each school are removed with the school fixed effects.

[35] Robust standard errors in Tables VII–IX are clustered at the school level to correct for general autocorrelations among the errors across cohorts of students attending the same school; for a discussion of the issue in a related context, see Bertrand, Duflo, and Mullainathan (2004).

TABLE VII

EFFECTS OF TEACHER AND SCHOOL CHARACTERISTICS ON 4TH–7TH GRADE GAINS IN MATHEMATICS AND READING TEST SCORES (ROBUST STANDARD ERRORS IN PARENTHESES; $n = 1{,}336{,}903$ FOR MATHEMATICS AND 1,330,791 FOR READING)

	No Fixed Effects	Student and School Fixed Effects	Student and School-by-Year Fixed Effects	Student, School-by-Grade and School-by-Year Fixed Effects
1. Mathematics				
Class size				
4th grade	−0.0049	−0.0106	−0.0107	n.a.
	(0.0023)	(0.0040)	(0.0037)	
5th grade	−0.0043	−0.0085	−0.0081	−0.0055
	(0.0010)	(0.0017)	(0.0024)	(0.0018)
6th grade	−0.0014	−0.0037	−0.0041	−0.0027
	(0.0010)	(0.0017)	(0.0020)	(0.0013)
7th grade	0.0002	0.0025	0.0032	0.0011
	(0.0009)	(0.0020)	(0.0024)	(0.0023)
Experience				
Proportion	−0.085	−0.103	−0.128	−0.073
0 years	(0.012)	(0.021)	(0.028)	(0.023)
Proportion	−0.043	−0.066	−0.055	−0.002
1 year	(0.013)	(0.022)	(0.028)	(0.023)
Proportion	−0.018	−0.045	−0.055	−.002
2 years	(0.013)	(0.021)	(0.030)	(0.022)
Proportion	−0.012	−0.031	−0.030	−0.017
3–5 years	(0.010)	(0.018)	(0.022)	(0.018)
Education				
Proportion with	−0.025	−0.018	−0.023	−0.021
graduate degree	(0.009)	(0.017)	(0.021)	(0.020)

6.2.1. *Class size*

The results reveal statistically significant effects of class size on both mathematics and reading achievement gains, but the impact declines markedly as students progress through school and tends to be smaller and less significant in reading than in mathematics. The discussion concentrates on the model that removes school-by-year fixed effects, because 4th grade estimates cannot be produced for models that contain school-by-grade fixed effects with only the single available 4th grade cohort.

The estimated effects of class size are quite similar quantitatively and qualitatively across specifications that include student and either school or school-by-year fixed effects.[36] Both the 4th and 5th grade class size coefficients are

[36]However, the addition of school-by-grade fixed effects substantially reduces the magnitudes and significance levels of estimates in mathematics though not in reading. Nevertheless, class size continues to exert a significant effect on mathematics achievement in grades 5 and 6. It is

TABLE VII—CONTINUED

	No Fixed Effects	Student and School Fixed Effects	Student and School-by-Year Fixed Effects	Student, School-by-Grade and School-by-Year Fixed Effects
2. Reading				
Class size				
4th grade	−0.0031	−0.0090	−0.0092	n.a.
	(0.0017)	(0.0031)	(0.0029)	
5th grade	0.0000	−0.0033	−0.0032	−0.0043
	(0.0007)	(0.0012)	(0.0018)	(0.0016)
6th grade	0.0021	0.0000	−0.0003	−0.0021
	(0.0009)	(0.0013)	(0.0019)	(0.0013)
7th grade	−0.0046	−0.0022	−0.0028	−0.0013
	(0.0008)	(0.0017)	(0.0024)	(0.0020)
Experience				
Proportion	−0.041	−0.045	−0.064	−0.026
0 years	(0.010)	(0.019)	(0.023)	(0.021)
Proportion	−0.037	−0.042	−0.070	−0.002
1 year	(0.010)	(0.018)	(0.023)	(0.020)
Proportion	−0.004	−0.006	−0.018	0.002
2 years	(0.010)	(0.019)	(0.025)	(0.020)
Proportion	0.001	0.014	0.002	0.018
3–5 years	(0.009)	(0.015)	(0.020)	(0.017)
Education				
Proportion with	−0.014	−0.004	0.001	0.010
graduate degree	(0.007)	(0.014)	(0.018)	(0.017)

Note: All specifications include a full set of grade-by-year dummies and indicators for subsidized lunch eligibility and a change of school prior to or during year. Robust standard errors in Tables VII–IX are clustered at the school level to correct for general autocorrelations among the errors across cohorts of students attending the same school; for a discussion of the issue in a related context, see Bertrand, Duflo, and Mullainathan (2004).

highly significant in both subjects, though the magnitude of the 5th grade effect is roughly three-fourths as large as that for 4th grade in mathematics and less than half as large in reading. The 6th grade effects are quite small, and by 7th grade class size appears to have little systematic effect on achievement. We discuss the magnitude of these estimates below. Note that the very large samples permit the precise estimation of quite small effects of less than 0.004 standard deviations.

The pattern of estimated class size effects also reveals the importance of controlling for student fixed effects. The inclusion of student fixed effects

not possible to know for certain the extent to which change with the addition of school-by-grade fixed effects results from the elimination of further biases as opposed to the exacerbation of any problems with measurement error.

TABLE VIII

EFFECTS OF CLASS SIZE ON TEST SCORE GAINS, BY FAMILY INCOME
(ROBUST STANDARD ERRORS IN PARENTHESES)

	Mathematics		Reading	
	Disadvantaged Students	Not Disadvantaged Students	Disadvantaged Students	Not Disadvantaged Students
Class size				
4th grade	−0.0118	−0.0103	−0.0111	−0.0087
	(0.0038)	(0.0037)	(0.0030)	(0.0029)
5th grade	−0.0077	−0.0079	−0.0027	−0.0033
	(0.0025)	(0.0024)	(0.0019)	(0.0018)
6th grade	−0.0044	−0.0040	−0.0022	−0.0007
	(0.0021)	(0.0020)	(0.0019)	(0.0017)
7th grade	0.0036	0.0031	0.0012	−0.0037
	(0.0026)	(0.0024)	(0.0023)	(0.0022)

Note: Estimates come from a single mathematics regression and a single reading regression. The models include student and school-by-year fixed effects, separate class size, and teacher experience variables for students eligible for a subsidized lunch (disadvantaged) and those not eligible during a given school year, proportion of teachers with a graduate degree, full sets of grade-by-year dummies, and indicators for subsidized lunch eligibility and a change of school prior to or during year.

triples the 4th grade coefficient and more than doubles the coefficient for 5th grade.[37]

An important and often studied question is whether lower income students receive larger benefits from class size reduction. In order to examine this claim we relaxed the restriction that class size effects were the same by income (measured by subsidized lunch eligibility). The results in Table VIII generally do not support the belief that class size effects are substantially larger for disadvantaged (subsidized lunch eligible) students. Class size effects are roughly 20 percent larger for disadvantaged students in 4th grade but actually smaller in 5th grade. Both the grade pattern and the comparable mathematics and reading results are very similar to the results in Table VII.

One potential perspective on these estimates comes from Project STAR, the random assignment experiment in class size reduction conducted in Tennessee (Word et al. (1990)).[38] While these experimental results are not directly comparable because they consider just grades K to 3, they indicate that a reduction

[37]The progressively more stringent estimates found across the columns does introduce some instability in the estimates, particularly in the final column. The smaller though still significant coefficients in the full fixed effects model for mathematics are consistent with the possibility that the school-by-grade and school-by-year fixed effects together aggravate problems associated with measurement error, but the results for reading go in the opposite direction.

[38]Project STAR randomly assigned a large group of kindergarten students to regular sized classes (22–25 students), regular sized classes with an aide, or small classes (13–17 students). It was designed to follow these students through grade 3, but there were significant attrition problems and subsequent additions of students to the experiment. Achievement tests were given

of eight students per class yields kindergarten achievement gains in math and reading of 0.17 standard deviations, which is roughly 60 percent larger than our 4th grade result for mathematics and reading. However, the deeper inconsistency that cannot be resolved here is that the experimental results indicate that virtually all of the achievement gain in STAR is associated with the first year in a small class—generally kindergarten or 1st grade—and not subsequent small class treatments (Krueger (1999)), while we find that smaller classes still have an effect in 4th and 5th grade.

The STAR experiment also reveals very large variation in student performance across individual classrooms. Specifically, all randomization occurred within each experimental school, and students in the large classes outperformed schoolmates in smaller classes in almost half of the schools (Hanushek (1999b)). This experimental finding is consistent with the conclusions here that differences in teacher quality within schools are quite large.

The school-by-year fixed effect estimates in column 3 of Table VII provide the basis for a simple comparison of policy alternatives. While it is difficult to estimate the cost of improving teacher quality, our lower bound estimates of the variation in quality found just within schools indicate that one standard deviation in quality is worth at least 0.11 standard deviations higher annual growth in mathematics achievement and 0.095 standard deviations higher annual growth in reading in elementary school. This magnitude of change is equivalent to a class size reduction of approximately ten students in 4th grade and thirteen or more students in 5th grade, and an implausibly large number in 6th grade. In 7th grade there appears to be no significant benefit from smaller classes in mathematics, while in reading neither class size nor teacher quality appears to exert a substantial effect on achievement. Note that these comparisons assume both no accompanying changes in teacher quality and linearity in class size effects, the latter of which appears reasonable based on semiparametric estimates for class sizes between 10 and 35 students (results not reported).

6.2.2. *Teacher characteristics*

The results for teacher experience generally support the notion that beginning teachers and to a lesser extent second and third year teachers in mathematics perform significantly worse than more experienced teachers. There may be some additional gains to experience in the subsequent year or two, but the estimated benefits are small and not statistically significant in both mathematics and reading in any of the fixed effect specifications. Similar to the case for class size, the results in the full fixed effect model in column 4 are much weaker

at the end of each grade, and a comparison showed that students in small classes outperformed those in regular classes in their first experimental year (K or 1) but that no additional gains were made. See Hanushek (1999b) and Krueger (1999).

than in the other fixed effects models, consistent with the view that multiple fixed effects can exacerbate problems with measurement error. The addition of school-by-grade fixed effects reduces the magnitude of all coefficients, and only the estimated effect of proportion of new teachers on math achievement gain is significant.

Importantly, the teacher experience effect conceptually combines two very distinct phenomena. First, new teachers may need to go through an adjustment period where they learn the craft of teaching along with adjusting to the other aspects of an initial job. Second, a number of the early teachers discover that they are not well matched for teaching and subsequently leave the profession within the first few years. Between entry and the end of two years, 18 percent of teachers will leave the Texas public schools, and another 6 percent will switch districts (Hanushek, Kain, and Rivkin (2004b)). The estimated parameters in Table VII combine the effects of on-the-job learning and of selective exit and mobility.

Table IX presents the basic estimates of first year teaching on achievement (with individual and school fixed effects) for samples that exclude those who immediately leave teaching or switch schools. The close similarity of the estimates across the samples compared to those in Table VII for both mathematics and reading indicates that on-the-job learning is the dominant element of the experience effect. Importantly, these results also suggest that the average quality of those who quit teaching after one year is similar to the average quality of those who remain, providing additional support for the validity of the estimates of the variance in teacher quality.

TABLE IX

EFFECTS OF PROPORTION OF TEACHERS WITH ZERO YEARS OF EXPERIENCE ON MATHEMATICS AND READING TEST SCORE GAINS, BY NEW TEACHER TRANSITIONS (ROBUST STANDARD ERRORS IN PARENTHESES)

Outcome Measure	Excluding Teachers Who Exit Teaching or Switch Schools	Excluding Teachers Who Exit Teaching	All Teachers
1. Mathematics			
Proportion of teachers with 0 years experience	−0.105	−0.114	−0.103
	(0.030)	(0.028)	(0.021)
Observations	[1,185,329]	[1,210,155]	[1,336,903]
2. Reading			
Proportion of teachers with 0 years experience	−0.040	−0.040	−0.045
	(0.024)	(0.023)	(0.019)
Observations	[1,181,611]	[1,206,139]	[1,330,791]

Note: Estimates come from a model that includes student and school fixed effects. Specifications also include the percentage of teachers with a graduate degree, full sets of class size variables, and grade-by-year dummies and indicators for subsidized lunch eligibility and a change of school prior to or during year.

Finally, consistent with previous work, there is little or no evidence that a master's degree raises the quality of teaching. All estimates are small (or negative) and statistically insignificant.

7. CONCLUSIONS

Prior investigations of school and teacher effects have raised as many questions as they have answered, in large part because of the difficulties introduced by the endogeneity of school and classroom selection and in part because of the failure of observable teacher characteristics to explain much of the variation in student performance. The models and data used in this paper permit us to draw a number of sharp conclusions about public elementary education and to provide clear answers for the questions raised in the Introduction.

(i) *Teachers and therefore schools matter importantly for student achievement.* The issue of whether or not there is significant variation in school quality has lingered, quite inappropriately, since the original Coleman Report. This analysis identifies large differences in the quality of instruction in a way that rules out the possibility that the observed differences are driven by family factors.

The Coleman Report also popularized the issue of whether family influences are "more important" than school influences. This is not the relevant question for policy, which should focus on whether the benefits produced by any intervention justify the costs. Though our analysis does not consider the costs of raising teacher quality, the estimated variation in the quality of instruction clearly reveals an important role for schools and teachers in promoting economic and social equality. Even if none of the between-school variation in achievement is attributed to schools or teachers, it is clear that school policy can be an important tool for raising the achievement of low income students and that a succession of good teachers could, by our estimates, go a long way toward closing existing achievement gaps across income groups. At the very least, more must be known about the feasible means of providing such consistently high quality teachers.

(ii) *Achievement gains are systematically related to observable teacher and school characteristics, but the effects are generally small and concentrated among younger students.* This analysis used a fixed effects approach to identify the causal relationship between achievement and key school resources. Four major conclusions emerge from this work.

- Similar to most past research, we find absolutely no evidence that having a master's degree improves teacher skills.
- There appear to be important gains in teaching quality in the first year of experience and smaller gains over the next few career years. However, there is little evidence that improvements continue after the first three years.
- Class size appears to have modest but statistically significant effects on mathematics and reading achievement growth that decline as students progress through school.

- Any differences in school resource effects by family income are small.

Partially consistent with recent experimental and statistical efforts to identify class size effects, we find that lowering class size has a positive effect on mathematics and reading achievement, though the magnitude of the effect is small, particularly following 5th grade. The costs of class size reduction have not been well estimated, but they are likely to exceed the proportional increase in the number of teachers needed to staff the smaller classes. First, class size reduction almost certainly leads to more support expenditure, increased building requirements, and the like. Second, and more directly relevant to this discussion, it is highly unlikely that the supply of teacher quality is perfectly elastic, so that expansion of the teacher work force, at least in the short run, is likely to lead either to increased salary demands or a reduction in teacher quality. Moreover, the potential tradeoff between teacher quality and class size is probably most acute in difficult to staff schools serving largely disadvantaged student populations (Hanushek (1999a), Jepsen and Rivkin (2002)).

(iii) *The disjuncture between estimates of the variation of teacher quality and the explanatory power of measured teacher characteristics creates a clear dilemma for policy makers.* Though it is tempting to tighten standards for teachers in an effort to raise quality, the results in this paper and elsewhere raise serious doubts that more restrictive certification standards, education levels, etc. will succeed in raising the quality of instruction. Rather the substantial differences in quality among those with similar observable backgrounds highlight the importance of effective hiring, firing, mentoring, and promotion practices. Research shows that principals can, when asked, separate teachers on the basis of quality (Murnane (1975), Armor et al. (1976)), but the substantial variation documented in this paper strongly suggests that personnel practices in the Texas public schools are very imperfect.

One dimension of policy does, nonetheless, deserve special attention. Economically disadvantaged students systematically achieve less than more advantaged students, on average falling some 0.6 standard deviations behind.[39] While we find little reason to believe that school resources have a larger impact on disadvantaged students, we do know that low income and minority students face higher teacher turnover and tend to be taught more frequently by beginning teachers (Hanushek, Kain, and Rivkin (2004b)). Because beginning teachers, regardless of their ultimate abilities, tend to perform more poorly, policies should be developed to both keep more senior teachers in the classrooms of disadvantaged students and to mitigate the impact of inexperience. These may include improved mentoring of new teachers and policies designed specifically to cut down teacher turnover. Of course, it goes without saying that

[39]The measure of family income is eligibility for a free or reduced price school lunch. This measure, while quite commonly used because of its availability in administrative records, is an imprecise categorization of economic circumstances.

effective policies will pay particular attention to the substantial variation in teacher quality.

The desirability of specific policy changes remains quite speculative because of the limited experience with alternative organizational forms, incentives, and accountability policies. A very appealing though untested approach to raising teacher quality would move the focus away from the state legislatures and schools of education and toward principals and other administrators (Hanushek and Rivkin (2004)). In the presence of incentives such as expanded choice, school report cards, or other types of accountability systems, administrators would likely alter their behavior and personnel policies in ways that benefit students. In particular, there would likely be much more focus on student outcomes of interest. Not only would improved personnel policies likely raise the performance level of existing teachers, there is strong reason to believe that a closer link between rewards and performance would improve the stock of teachers. Of course inappropriate incentives likely lead to adverse outcomes, and it is imperative that schools learn from their mistakes and evolve toward more effective systems of school governance.

Dept. of Economics, Amherst College, Amherst, MA 01002, U.S.A.; sgrivkin@amherst.edu,

Hoover Institution, Stanford University, Stanford, CA 94305, U.S.A.; hanushek@stanford.edu; http://www.hanushek.net,

and

University of Texas at Dallas (deceased).

Manuscript received July, 2002; final revision received October, 2004.

APPENDIX A: THE EFFECT OF NONRANDOM TEACHER ATTRITION ON TURNOVER-BASED ESTIMATOR OF TEACHER QUALITY VARIATION

The estimator of teacher quality derived from equation (7) assumes that the error term (e) is uncorrelated with teacher turnover. If, however, there is systematic teacher attrition that varies by quality, the estimator may no longer be a lower bound but may in fact overestimate the variance in quality. This specifically would be the case if attrition is concentrated in the tails of the quality distribution. It is most natural to think of this as a problem of sample selection where teachers who depart have a different distribution in terms of quality than those who remain. Thus, schools with turnover would tend to have a different quality distribution for teachers.

The nature of the problem with selective attrition using our estimator is easiest to see in the simpler comparison of the squared difference in grade g gains for successive cohorts, although it would easily generalize to the full estimator. The subtraction of 5th grade average gain from 6th grade average gain for a cohort removes any student and school fixed effects (including overall hiring practices) but does not address problems related to nonrandom teacher departures.

TABLE A1

UNDERLYING DISTRIBUTION OF TEACHER QUALITY FOR NEW HIRES

Relative Teacher Quality (q)	Frequency: $f(q)$
−1	0.25
0	0.50
1	0.25

The potential impact of selective attrition is directly seen from a simple simulation using a trinomial quality distribution. Table A1 describes a distribution of new hires that has a variance of quality equal to 0.5. With this distribution of new hires, it is possible to simulate the estimator of school quality both with random departures and with systematic departures that differ across the distribution.

First, consider the turnover-based estimator of the variance in teacher quality when there are random departures. Table A2 begins with the distribution of teacher quality in Table A1 and then assumes that teachers leave randomly (and are replaced by a random selection of teachers according to the distribution in Table A1). Consequently there are nine possible transitions, three for each of the period 0 quality categories.

In this simple one grade example, the expected period 0/period 1 difference in quality is two times the variance in teacher quality (instead of four times the variance as derived in the full estimator that considers deviations across grades and cohorts). Table A2 shows that the estimator yields the true variation in quality when there is random hiring and departures.

Consider, however, the identical estimator with strongly nonrandom departures characterized by probabilities of departure of 0.5, 0.0, and 0.5 for the

TABLE A2

TRANSITION MATRIX AND VARIANCE ESTIMATE WITH RANDOM ATTRITION

Relative Teacher Quality (q_0) Period 1	Relative Teacher Quality (q_1) Period 2	Transition Frequency: $f(q_1, q_0)$	Squared Quality Difference $(q_1 - q_0)^2$
	−1	0.0625	0
−1	0	0.125	1
	+1	0.0625	4
	−1	0.125	1
0	0	0.250	0
	+1	0.125	1
	−1	0.0625	4
+1	0	0.125	1
	+1	0.0625	0

Notes: Weighted sum of squared differences = 1.0; estimated variance = 1/2 squared differences = 0.5.

TABLE A3

TRANSITION MATRIX AND VARIANCE ESTIMATE WITH NONRANDOM ATTRITION CONCENTRATED IN THE TAILS OF THE QUALITY DISTRIBUTION

Relative Teacher Quality (q_0) Period 1	Relative Teacher Quality (q_1) Period 2	Transition Frequency: $f(q_1, q_0)$	Squared Quality Difference $(q_1 - q_0)^2$
	−1	0.125	0
−1	0	0.250	1
	+1	0.125	4
	−1	0.0	1
0	0	0.0	0
	+1	0.0	1
	−1	0.125	4
+1	0	0.250	1
	+1	0.125	0

Notes: Weighted sum of squared differences = 1.5; estimated variance = 1/2 squared differences = 0.75.

three quality groups in Table A1. Table A3 describes the transition probabilities, sum of squared quality differences, and the simulated variance estimates. If departures were as concentrated in the tails of the distribution as they are in this example, our method would overstate the variance in teacher quality by 50 percent: 0.75 instead of 0.5. Note that this upward bias would also arise if all departures were concentrated in only one of the tails of the distribution.

In general, if attrition is weighted toward the tails of the quality distribution the turnover-based estimator will tend to overestimate the variance of quality, and the opposite will hold if attrition is concentrated in the center of the quality distribution.

APPENDIX B: TEXAS SCHOOL DATA

The data that are used in this paper come from the data development activity of the UTD Texas Schools Project of the University of Texas at Dallas; see Kain (2001). Working with the Texas Education Agency (TEA), this project has combined a number of different data sources to compile an extensive data set on schools, teachers, and students. Demographic information on students and teachers is taken from the PEIMS (Public Education Information Management System), which is TEA's statewide educational data base. Test score results and a limited amount of student demographic information are stored in a separate data base maintained by TEA and must be merged with the student data on the basis of unique student IDs. Data are compiled for all public school students in Texas, allowing us to use the universe of students in the analyses. In this paper all of the information on students comes from the test score data base, and we combine student information from the Texas Assessment of Academic Skills (TAAS) data base with teacher and school information contained in the PEIMs data base for three student cohorts: 3rd through 7th grade test

TABLE B1

VARIABLE MEANS AND STANDARD DEVIATIONS

				Teacher Characteristics			
	Math Test Score Gain	Reading Test Score Gain	Class Size	% with Graduate Degree	% 0 Years Experience	% 1 Year Experience	Observations
4th grade	−0.01	−0.02	19.5	23.7	6.1	5.9	143,314
	(0.70)	(0.73)	(2.3)	(24.3)	(12.4)	(12.5)	
5th grade	0.01	0.01	22.6	25.1	5.9	6.0	438,561
	(0.64)	(0.68)	(3.6)	(26.2)	(13.7)	(13.6)	
6th grade	0.02	0.02	22.1	24.5	7.4	6.9	455,438
	(0.61)	(0.68)	(3.9)	(27.4)	(16.6)	(15.7)	
7th grade	−0.02	−0.01	21.5	22.0	9.2	8.9	299,590
	(0.55)	(0.66)	(4.2)	(26.7)	(18.4)	(18.0)	

scores for one cohort (4th graders in 1995) and 4th through 7th grade test scores for the other two (4th graders in 1993 and 1994).[40]

Beginning in 1993, the Texas Assessment of Academic Skills (TAAS) was administered each spring to eligible students enrolled in grades 3 through 8. We focus on test results for mathematics and reading. The bottom one percent of test scores are trimmed from the sample in order to reduce measurement error. Participants in bilingual or special education programs are also excluded from the sample, because of the difficulty in measuring school and teacher characteristics for students who split time between regular classrooms and special programs.

Student data are merged with information on teachers using unique school identifiers. The personnel data provide information on all Texas public school teachers for each year. Experience and highest degree earned are reported, as are the class size, subject, grade, and population served for each class taught. Although the currently available data do not permit linking individual students with specific teachers, the available information is used to construct subject and grade average characteristics for teachers in regular classrooms.

In an effort to reduce problems associated with measurement error, a number of observations are excluded from the data set. The following paragraphs describe in detail the construction of the variables and the sample selection procedures.

Measurement error in the teacher characteristics is an important issue. In many cases reported teacher experience in one year does not correspond with reported teacher experience for other years. If the experience sequence is valid except for one or two years that do not follow from the others, we correct ex-

[40]Note that, while we have 3rd grade test information, our analysis begins at 4th grade because of the focus on achievement gains.

perience for those years. If experience data are inconsistent for all the years, if there are two consistent patterns, or if correction would impute negative years of experience, no corrections are made. In any case, no teachers are excluded from the final sample on the basis of inconsistent experience data, though the results are not sensitive to their inclusion, possibly because we used discreet experience categories.

The case of average class size is somewhat more complicated. Teachers were asked to report the average class size for each class they taught that was of a different size. Unfortunately, many teachers appear to have reported the total number of students taught per day. This becomes particularly problematic for schools that move from general to subject specific teachers. Consider a school with two 4th grade classes of twenty students in which the two teachers each teach all subjects. If the school switches to math and reading specialists for 5th grade and each teaches one subject for each class, they will report class sizes of forty if they report total number of students served. It will appear that class sizes doubled as students aged, when in fact they remain the same.

In order to reduce problems introduced by measurement, all reported class sizes that fall below ten or above twenty five in 4th grade (thirty five in higher grades) are set to missing prior to the computation of school averages for each grade. By statute, 4th grade classes are not supposed to exceed twenty two students, though some schools receive waivers to provide slightly larger classes. It is our understanding that very few elementary schools in Texas have actual class sizes in later grades that exceed thirty five students during this period. Estimates of class size effects increased in magnitude following these exclusions, suggesting that class size was measured with error for these schools.

Access to the administrative data on student performance is currently restricted by U.S. federal law. Further information on data access along with the specific variable definitions, data construction, and data that may currently be released are found in Rivkin, Hanushek, and Kain (2005).

APPENDIX C: ALTERNATIVE TEACHER QUALITY ESTIMATES

TABLE C1

TEACHER QUALITY STANDARD DEVIATION ESTIMATES CALCULATED FROM SQUARED DIFFERENCE IN QUALITY FOR PERIODS 0 AND 1, BASED ON OBSERVED DISTRIBUTIONS OF TEACHER QUALITY AND DEPARTURE RATES

Number of Teacher Quality Intervals	σ Assuming Random Departures	σ Assuming Empirical Distribution of Departures
20 (Table VI)	0.395	0.399
40	0.397	0.401
60	0.397	0.402
30 with tails	0.422	0.427

TABLE C2

EFFECT OF TEACHER TURNOVER ON THE DIVERGENCE OF GAINS IN MATHEMATICS AND READING TEST SCORES BETWEEN COHORTS FOR SCHOOLS WITH ONE TEACHER PER GRADE (STANDARD ERRORS IN PARENTHESES)

	No Fixed Effects	Individual and School Fixed Effects	Individual and School-by-Grade Fixed Effects
1. Mathematics			
Proportion different math teachers/number of teachers	0.124 (0.039)	0.117 (0.039)	0.042 (0.047)
2. Reading			
Proportion different English teachers/number of teachers	0.181 (0.037)	0.180 (0.049)	0.061 (0.042)

Notes: All equations include the inverse of the number of students, numbers of new principals and superintendents in the school during adjacent years, and a cohort dummy variable. Sample size is 294 for the mathematics and 300 for the reading specifications. Table III notes describe the estimation specifications.

REFERENCES

ANGRIST, J. D., AND V. LAVY (1999): "Using Maimondides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533–575.

ARMOR, D. J., P. CONRY-OSEGUERA, M. COX, N. KING, L. MCDONNELL, A. PASCAL, E. PAULY, AND G. ZELLMAN (1976): *Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools*. Santa Monica, CA: Rand Corp.

BERTRAND, M., E. DUFLO, AND S. MULLAINATHAN (2004): "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 114, 249–275.

BETTS, J. R. (1995): "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth," *Review of Economics and Statistics*, 77, 231–247.

BONESRØNNING, H. (2004): "The Determinants of Parental Effort in Education Production: Do Parents Respond to Changes in Class Size?" *Economics of Education Review*, 23, 1–9.

BOOZER, M. A., AND C. ROUSE (1995): "Intraschool Variation in Class Size: Patterns and Implications," Working Paper 5144, National Bureau of Economic Research, Cambridge, MA.

BURTLESS, G. (ED.) (1996): *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*. Washington, DC: Brookings.

CARD, D., AND A. B. KRUEGER (1992): "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100, 1–40.

COLEMAN, J. S., E. Q. CAMPBELL, C. J. HOBSON, J. MCPARTLAND, A. M. MOOD, F. D. WEINFELD, AND R. L. YORK (1966): *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

DEARDEN, L., J. FERRI, AND C. MEGHIR (2002): "The Effect of School Quality on Educational Attainment and Wages," *Review of Economics and Statistics*, 84, 1–20.

DOBBELSTEEN, S., J. LEVIN, AND H. OOSTERBEEK (2002): "The Causal Effect of Class Size on Scholastic Achievement: Distinguishing the Pure Class Size Effect from the Effect of Changes in Class Composition," *Oxford Bulletin of Economics and Statistics*, 64, 17–38.

DOLTON, P. J., AND W. VAN DER KLAAUW (1995): "Leaving Teaching in the UK: A Duration Analysis," *The Economic Journal*, 105, 431–444.

——— (1999): "The Turnover of Teachers: A Competing Risks Explanation," *Review of Economics and Statistics*, 81, 543–552.

DUSTMANN, C., N. RAJAH, AND A. VAN SOEST (2003): "Class Size, Education, and Wages," *Economic Journal*, 113, F99–120.

FEINSTEIN, L., AND J. SYMONS (1999): "Attainment in Secondary Schools," *Oxford Economic Papers*, 52, 300–321.

GREENWALD, R., L. V. HEDGES, AND R. D. LAINE (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*, 60, 280–288.

——— (1979): "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources*, 14, 351–388.

——— (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24, 1141–1177.

——— (1992): "The Trade-Off Between Child Quantity and Quality," *Journal of Political Economy*, 100, 84–117.

——— (1996): "A More Complete Picture of School Resource Policies," *Review of Educational Research*, 66, 397–409.

——— (1999a): "The Evidence on Class Size," in *Earning and Learning: How Schools Matter*, ed. by S. E. Mayer and P. E. Peterson. Washington, DC: Brookings Institution, 131–168.

——— (1999b): "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects," *Educational Evaluation and Policy Analysis*, 21, 143–163.

——— (2003): "The Failure of Input-Based Schooling Policies," *Economic Journal*, 113, F64–F98.

HANUSHEK, E. A., AND J. F. KAIN (1972): "On the Value of 'Equality of Educational Opportunity' as a Guide to Public Policy," in *On Equality of Educational Opportunity*, ed. by F. Mosteller and D. P. Moynihan. New York: Random House, 116–145.

HANUSHEK, E. A., J. F. KAIN, D. M. O'BRIEN, AND S. G. RIVKIN (2005): "The Market for Teacher Quality," Working Paper, National Bureau of Economic Research, Cambridge, MA.

HANUSHEK, E. A., J. F. KAIN, AND S. G. RIVKIN (2002): "Inferring Program Effects for Specialized Populations: Does Special Education Raise Achievement for Students with Disabilities?" *Review of Economics and Statistics*, 84, 584–599.

——— (2004a): "Disruption versus Tiebout Improvement: The Costs and Benefits of Switching Schools," *Journal of Public Economics*, 88/9–10, 1721–1746.

——— (2004b): "Why Public Schools Lose Teachers," *Journal of Human Resources*, 39, 326–354.

HANUSHEK, E. A., AND S. G. RIVKIN (2004): "How to Improve the Supply of High Quality Teachers," in *Brookings Papers on Education Policy 2004*, ed. by D. Ravitch. Washington, DC: Brookings Institution Press, 7–25.

HECKMAN, J. J., A. LAYNE-FARRAR, AND P. TODD (1996): "Human Capital Pricing Equations with an Application to Estimating the Effect of Schooling Quality on Earnings," *Review of Economics and Statistics*, 78, 562–610.

HOXBY, C. M. (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, 115, 1239–1285.

INGERSOLL, R. M. (2001): "Teacher Turnover and Teacher Shortages: An Organizational Analysis," *American Educational Research Journal*, 38, 499–534.

JEPSEN, C., AND S. G. RIVKIN (2002): "What Is the Trade-Off Between Smaller Classes and Teacher Quality?" National Bureau of Economic Research, Cambridge, MA.

KAIN, J. F. (2001): "The UTD Texas Schools Microdata Panel (TSMP): Its History, Use and Ways to Improve State Collection of Public School Data," Paper Prepared for The Secretary's Forum on Research and Value-Added Assessment Data, U.S. Department of Education; *http://utdallas.edu/research/tsp/index.htm.*

KAIN, J. F., AND D. M. O'BRIEN (1998): "A Longitudinal Assessment of Reading Achievement: Evidence from the Harvard/UTD Texas Schools Project," University of Texas at Dallas, Dallas, TX.

KRUEGER, A. B. (1999): "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, 497–532.

LEVAČIĆ, R., AND A. VIGNOLES (2002): "Researching the Links Between School Resources and Student Outcomes in the UK: A Review of Issues and Evidence," *Education Economics*, 10, 313–331.

MURNANE, R. J. (1975): *Impact of School Resources on the Learning of Inner City Children*. Cambridge, MA: Ballinger.

——— (1984): "Selection and Survival in the Teacher Labor Market," *Review of Economics and Statistics*, 66, 513–518.

MURNANE, R. J., AND R. OLSEN (1989): "The Effects of Salaries and Opportunity Costs on Length of Stay in Teaching: Evidence from Michigan," *Review of Economics and Statistics*, 71, 347–352.

MURNANE, R. J., AND B. PHILLIPS (1981): "What Do Effective Teachers of Inner-City Children Have in Common?" *Social Science Research*, 10, 83–100.

MURNANE, R. J., J. D. SINGER, J. B. WILLETT, J. J. KEMPLE, AND R. J. OLSEN (1991): *Who Will Teach? Policies That Matter*. Cambridge, MA: Harvard University Press.

RIVKIN, S. G., E. A. HANUSHEK, AND J. F. KAIN (2005): "Variable Definitions, Data, and Programs for 'Teachers, Students, and Academic Achievement'," *Econometrica Supplementary Material*, 73, 2, *www.econometricsociety.org/ecta/supmat/4139data.pdf*.

ROBERTSON, D., AND J. SYMONS (2003): "Do Peer Groups Matter? Peer Group versus Schooling Effects on Academic Attainment," *Economica*, 70, 31–53.

STINEBRICKNER, T. R. (2002): "An Analysis of Occupational Change and Departure from the Labor Force," *Journal of Human Resources*, 37, 192–216.

SUMMERS, A. A., AND B. L. WOLFE (1977): "Do Schools Make a Difference?" *American Economic Review*, 67, 639–652.

THE TEACHING COMMISSION (2004): *Teaching at Risk: A Call to Action*. New York, NY: The Teaching Commission.

TIEBOUT, C. M. (1956): "A Pure Theory of Local Expenditures," *Journal of Political Economy*, 64, 416–424.

WOESSMANN, L. (2004): "Educational Production in Europe," Paper Presented at 40th Meeting of *Economic Policy* in Amsterdam Ifo Institute for Economic Research at the University of Munich.

WOESSMANN, L., AND M. R. WEST (forthcoming): "Class-Size Effects in School Systems Around the World: Evidence from Between-Grade Variation in TIMSS," *European Economic Review*, forthcoming.

WORD, E., J. JOHNSTON, H. P. BAIN, B. D. FULTON, J. B. ZAHARIES, M. N. LINTZ, C. M. ACHILLES, J. FOLGER, AND C. BREDA (1990): *Student/Teacher Achievement Ratio (STAR), Tennessee's K-3 Class Size Study: Final Summary Report, 1985–1990*. Nashville, TN: Tennessee State Department of Education.

EXHIBIT 5
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

Teachers and Student Achievement in the Chicago Public High Schools

Daniel Aaronson, *Federal Reserve Bank of Chicago*

Lisa Barrow, *Federal Reserve Bank of Chicago*

William Sander, *DePaul University*

We estimate the importance of teachers in Chicago public high schools using matched student-teacher administrative data. A one standard deviation, one semester improvement in math teacher quality raises student math scores by 0.13 grade equivalents or, over 1 year, roughly one-fifth of average yearly gains. Estimates are relatively stable over time, reasonably impervious to a variety of conditioning variables, and do not appear to be driven by classroom sorting or selective score reporting. Also, teacher quality is particularly important for lower-ability students. Finally, traditional human capital measures—including those determining compensation—explain little of the variation in estimated quality.

We thank the Chicago Public Schools and the Consortium on Chicago School Research at the University of Chicago for making the data available to us. We are particularly grateful to John Easton and Jenny Nagaoka for their help in putting together the data and answering our many follow-up questions. We thank Joe Altonji, Kristin Butcher, Dave Card, Rajeev Dehejia, Tom DiCiccio, Eric French, Brian Jacob, Jeff Kling, Steve Rivkin, Doug Staiger, Dan Sullivan, Chris Taber, and seminar participants at many universities and conferences for helpful comments and discussions. The views expressed in this article are ours and are not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System. Contact the corresponding author, Lisa Barrow, at lbarrow@frbchi.org.

[*Journal of Labor Economics,* 2007, vol. 25, no. 1]
© 2007 by The University of Chicago. All rights reserved.
0734-306X/2007/2501-0004$10.00

I. Introduction

The Coleman Report (Coleman et al. 1966) broke new ground in the estimation of education production functions, concluding that family background and peers were more important than schools and teachers in educational outcomes such as test scores and graduation rates. While research since Coleman supports the influence of family background, substantiation of the importance of other factors, particularly schools and teachers, has evolved slowly with the release of better data. Today, most researchers agree that schools and teachers matter.[1] However, how much they matter, the degree to which they vary across subpopulations, how robust quality rankings are to specification choices, and whether measurable characteristics such as teacher education and experience affect student educational outcomes continue to be of considerable research and policy interest.

In this study, we use administrative data from the Chicago public high schools to estimate the importance of teachers on student mathematics test score gains and then relate our measures of individual teacher effectiveness to observable characteristics of the instructors. Our measure of teacher quality is the effect on ninth-grade math scores of a semester of instruction with a given teacher, controlling for eighth-grade math scores and student characteristics. Our data provide us with a key advantage in generating this estimate: the ability to link teachers and students in specific classrooms. In contrast, many other studies can only match students to the average teacher in a grade or school. In addition, because teachers are observed in multiple classroom settings, our teacher effect estimates are less likely to be driven by idiosyncratic class effects. Finally, the administrative teacher records allow us to separate the effects of observed teacher characteristics from unobserved aspects of teacher quality.

Consistent with earlier studies, we find that teachers are important inputs in ninth-grade math achievement. Namely, after controlling for initial ability (as measured by test scores) and other student characteristics, teacher effects are statistically important in explaining ninth-grade math test score achievement, and the variation in teacher effect estimates is large

[1] Literature reviews include Greenwald, Hedges, and Laine (1996) and Hanushek (1996, 1997, 2002). A brief sampling of other work on teacher effects includes Murnane (1975), Goldhaber and Brewer (1997), Angrist and Lavy (2001), Jepsen and Rivkin (2002), Rivers and Sanders (2002), Jacob and Lefgren (2004), Rockoff (2004), Kane and Staiger (2005), Rivkin, Hanushek, and Kain (2005), and Kane, Rockoff, and Staiger (2006). The earliest studies on teacher quality were hampered by data availability and thus often relied on state- or school-level variation. Aggregation and measurement error compounded by proxies such as student-teacher ratios and average teacher experience can introduce significant bias. More recent studies, such as Rockoff (2004), Kane and Staiger (2005), Rivkin et al. (2005), and Kane et al. (2006), use administrative data like ours to minimize these concerns.

enough such that the expected difference in math achievement between having an average teacher and one that is one standard deviation above average is educationally important. However, a certain degree of caution must be exercised in estimating teacher quality using teacher fixed effects as biases related to measurement, particularly due to small populations of students used to identify certain teachers, can critically influence results. Sampling variation overstates our measures of teacher quality dispersion by amounts roughly similar to Kane and Staiger's (2002, 2005) evaluations of North Carolina schools and Los Angeles teachers. Correcting for sampling error, we find that the standard deviation in teacher quality in the Chicago public high schools is at least 0.13 grade equivalents per semester. Thus, over two semesters, a one standard deviation improvement in math teacher quality translates into an increase in math achievement equal to 22% of the average annual gain. This estimate is a bit higher than, but statistically indistinguishable from, those reported in Rockoff (2004) and Rivkin et al. (2005).[2]

Furthermore, we show that our results are unlikely to be driven by classroom sorting or selective use of test scores and, perhaps most importantly, the individual teacher ratings are relatively stable over time and reasonably impervious to a wide variety of conditioning variables. The latter result suggests that test score value-added measures for teacher productivity are not overly sensitive to reasonable statistical modeling decisions, and thus incentive schemes in teacher accountability systems that rely on similar estimates of productivity are not necessarily weakened by large measurement error in teacher productivity.

We also show how estimates vary by initial (eighth-grade) test scores, race, and sex and find that the biggest impact of a higher quality teacher, relative to the mean gain of that group, is among African American students and those with low or middle range eighth-grade test scores. We find no difference between boys and girls.

Finally, the vast majority of the variation in teacher effects is unexplained by easily observable teacher characteristics, including those used for compensation. While some teacher attributes are consistently related to our quality measure, together they explain at most 10% of the total variation in estimated teacher quality. Most troubling, the variables that determine compensation in Chicago—tenure, advanced degrees, and teaching certifications—explain roughly 1% of the total variation in es-

[2] Rivkin et al.'s (2005) lower bound estimates suggest that a one standard deviation increase in teacher quality increases student achievement by at least 0.11 standard deviations. Rockoff (2004) reports a 0.1 standard deviation gain from a one standard deviation increase in teacher quality from two New Jersey suburban school districts. In our results, a one standard deviation increase in teacher quality over a full year implies about a 0.15 standard deviation increase in math test score gains.

timated teacher quality. These results highlight the lack of a close relationship between teacher pay and productivity and the difficulty in developing compensation schedules that reward teachers for good work based solely on certifications, degrees, and other standard administrative data. That is not to say such schemes are not viable. Here, the economically and statistically important persistence of teacher quality over time should be underscored. By using past performance, administrators can predict teacher quality. Of course, such a history might not exist when recruiting, especially for rookie teachers, or may be overwhelmed by sampling variation for new hires, a key hurdle in prescribing recruitment, retention, and compensation strategies at the beginning of the work cycle. Nevertheless, there is clearly scope for using test score data among other evaluation tools for tenure, compensation, and classroom organization decisions.

While our study focuses on only one school district over a 3-year period, this district serves a large population of minority and lower income students, typical of many large urban districts in the United States. Fifty-five percent of ninth graders in the Chicago public schools are African American, 31% are Hispanic, and roughly 80% are eligible for free or reduced-price school lunch. Similarly, New York City, Los Angeles Unified, Houston Independent School District, and Philadelphia City serve student populations that are 80%–90% nonwhite and roughly 70%–80% eligible for free or reduced-price school lunch (U.S. Department of Education 2003). Therefore, on these dimensions Chicago is quite representative of the school systems that generate the most concern in education policy discussions.

II. Background and Data

The unique detail and scope of our data are major strengths of this study. Upon agreement with the Chicago Public Schools (CPS), the Consortium on Chicago School Research at the University of Chicago provided us with administrative records from the city's public high schools. These records include all students enrolled and teachers working in 88 CPS high schools from 1996–97 to 1998–99.[3] We concentrate on the performance of ninth graders in this article.

The key advantage to using administrative records is being able to work with the population of students, a trait of several other recent studies, including Rockoff (2004), Kane and Staiger (2005), Rivkin et al. (2005), and Kane et al. (2006). Apart from offering a large sample of urban schoolchildren, the CPS administrative records provide several other useful fea-

[3] Of the 88 schools, six are so small that they do not meet criteria on sample sizes that we describe below. These schools are generally more specialized, serving students who have not succeeded in the regular school programs.

tures that rarely appear together in other studies. First, this is the first study that we are aware of that examines high school teachers. Clearly, it is important to understand teacher effects at all points in the education process. Studying high schools has the additional advantage that classrooms are subject specific, and our data provide enough school scheduling detail to construct actual classrooms. Thus, we can examine student-teacher matches at a level that plausibly corresponds with what we think of as a teacher effect. This allows us to isolate the impact of math teachers on math achievement gains. However, we can go even further by, say, looking at the impact of English teachers on math gains. In this study, we report such exercises as robustness checks, but data like these offer some potential for exploring externalities or complementarities between teachers.

The teacher records also include specifics about human capital and demographics. These data allow us to decompose the teacher effect variation into shares driven by unobservable and observable factors, including those on which compensation is based. Finally, the student and teacher records are longitudinal. This has several advantages. Although our data are limited to high school students, they include a history of pre–high school test scores that can be used as controls for past (latent) inputs. Furthermore, each teacher is evaluated based on multiple classrooms over (potentially) multiple years, thus mitigating the influence of unobserved idiosyncratic class effects.

A. Student Records

There are three general components of the student data: test scores, school and scheduling variables, and family and student background measures. Like most administrative data sets, the latter is somewhat limited. Table 1 includes descriptive statistics for some of the variables available, including sex, race, age, eligibility for the free or reduced school lunch program, and guardian (mom, dad, grandparent, etc.). Residential location is also provided, allowing us to incorporate census tract information on education, income, and house values. We concentrate our discussion below on the test score and scheduling measures that are less standard.

1. *Test Scores*

In order to measure student achievement, we rely on student test scores from two standardized tests administered by the Chicago Public Schools—the Iowa Test of Basic Skills (ITBS) administered in the spring of grades 3–8 and the Test of Achievement and Proficiency (TAP) administered during the spring for grades 9 and 11.[4] We limit the study to

[4] TAP testing was mandatory for grades 9 and 11 through 1998. The year 1999 was a transition year in which ninth, tenth, and eleventh graders were tested. Starting in 2000, TAP testing is mandatory for grades 9 and 10.

Table 1
Descriptive Statistics for the Student Data

	All Students (1)		Students with Eighth- and Ninth-Grade Math Test Scores (2)		Students with Eighth- and Ninth-Grade Math Test Scores 1 Year Apart (3)	
Sample size:						
Total	84,154		64,423		52,957	
1997	29,301		21,992		17,941	
1998	27,340		20,905		16,936	
1999	27,513		21,526		18,080	
	Mean	SD	Mean	SD	Mean	SD
Test scores (grade equivalents):						
Math, ninth grade	9.07	2.74	9.05	2.71	9.21	2.64
Math, eighth grade	7.75	1.55	7.90	1.50	8.07	1.41
Math change, eighth to ninth grade	1.15	1.89	1.15	1.89	1.14	1.75
Reading comprehension, ninth grade	8.50	2.94	8.50	2.89	8.63	2.88
Reading comprehension, eighth grade	7.64	1.94	7.82	1.88	8.01	1.80
Reading change, eighth to ninth grade	.66	2.02	.67	2.02	.62	1.95
Demographics:						
Age	14.8	.8	14.7	.8	14.6	.7
Female	.497	.500	.511	.500	.522	.500
Asian	.035	.184	.033	.179	.036	.185
African American	.549	.498	.570	.495	.562	.496
Hispanic	.311	.463	.304	.460	.307	.461
Native American	.002	.047	.002	.046	.002	.046
Eligible for free school lunch	.703	.457	.721	.448	.728	.445
Eligible for reduced-price school lunch	.091	.288	.097	.295	.103	.304
Legal guardian:						
Dad	.241	.428	.244	.429	.253	.435
Mom	.620	.485	.626	.484	.619	.486
Nonrelative	.041	.197	.039	.195	.037	.189
Other relative	.038	.191	.034	.182	.032	.177
Stepparent	.002	.050	.002	.047	.002	.046
Schooling:						
Take algebra	.825	.380	.865	.342	.950	.217
Take geometry	.101	.302	.092	.290	.022	.145
Take computer science	.003	.054	.003	.057	.003	.057
Take calculus	.0001	.011	.0001	.010	.0001	.008
Fraction honors math classes	.081	.269	.093	.286	.101	.297
Fraction regular math classes	.824	.360	.827	.356	.820	.361
Fraction essential math classes	.032	.172	.029	.163	.032	.172
Fraction basic math classes	.001	.036	.001	.031	.001	.034
Fraction special education math classes	.014	.114	.009	.093	.009	.093
Fraction nonlevel math classes	.006	.057	.005	.054	.006	.057
Fraction level missing math classes	.042	.166	.036	.146	.030	.125
Fraction of math grades that are A	.083	.256	.085	.257	.093	.267
Fraction of math grades that are B	.130	.297	.138	.304	.151	.313
Fraction of math grades that are C	.201	.351	.218	.359	.232	.364
Fraction of math grades that are D	.233	.371	.250	.378	.252	.374
Fraction of math grades that are F	.311	.430	.272	.410	.241	.389
Fraction of math grades missing	.042	.166	.036	.146	.030	.125

Table 1 (*Continued*)

	All Students (1)		Students with Eighth- and Ninth-Grade Math Test Scores (2)		Students with Eighth- and Ninth-Grade Math Test Scores 1 Year Apart (3)	
Number of math/computer science classes taken in ninth grade	2.1	.4	2.1	.4	2.1	.4
Number of times in ninth grade	1.10	.31	1.08	.28	1.00	.00
Changed school within the year	.034	.180	.030	.170	.027	.163
Average class size among ninth-grade math classes	22.7	7.5	23.2	7.4	23.6	7.5
Cumulative GPA, spring	1.71	1.08	1.82	1.04	1.93	1.03
Average absences in ninth-grade math	13.9	16.7	11.6	13.7	9.9	11.7
Identified as disabled	.021	.143	.024	.154	.022	.147

NOTE.—The share of students disabled does not include students identified as learning disabled. Roughly 9% of CPS students in our estimation sample are identified as learning disabled.

ninth-grade students and primarily limit our analysis to math test scores. By limiting the study to ninth-grade students, we can also limit the sample to students with test scores from consecutive years in order to ensure that we associate math achievement with the student's teacher exposure in that same year. Although we also have information on reading test scores, we choose to focus on math achievement because the link between math teachers and math test scores is cleaner than for any single subject and reading scores. In addition, math test scores seem to have more, or are often assumed to have more, predictive power than reading scores for future productivity (see, e.g., Murnane et al. 1991; Grogger and Eide 1995; and Hanushek and Kimko 2000).

Multiple test scores are vital, as important family background measures, particularly income and parental education, are unavailable. While there are various ways to account for the cumulative effect of inputs that we cannot observe, we rely on a general form of the value-added model of education production in which we regress the ninth-grade test score on the variables of interest while controlling for initial achievement as measured by the previous year's eighth-grade test score.

We observe both eighth- and ninth-grade test scores for the majority of ninth-grade students, as shown in table 1. Scores are reported as grade equivalents, a national normalization that assigns grade levels to test score results in order to evaluate whether students have achieved the skills that are appropriate for their grade. For instance, a 9.7 implies that the student is performing at the level of a typical student in the seventh month of ninth grade. Unique student identifiers allow us to match the ninth-grade students to both their ninth-grade TAP score and their eighth-grade ITBS score.

Eighth- and ninth-grade test score data are reported for between 75% and 78% of the ninth-grade students in the CPS, yielding a potential sample of around 64,000 unique students over the 3-year period. Our sample drops to 53,000 when we exclude students without eighth- and ninth-grade test scores in consecutive school years and those with test score gains in the 1st and 99th percentiles.

Since the ninth-grade test is not a high stakes test for either students or teachers, it is less likely to elicit "cheating" in any form compared to the explicit teacher cheating uncovered in Jacob and Levitt (2003). In addition, by eliminating the outlier observations in terms of test score gains, we may drop some students for whom either the eighth- or ninth-grade test score is "too high" due to cheating. That said, there may be reasonable concern that missing test scores reflect some selection about which students take the tests or which test scores are reported.

Approximately 11% of ninth graders do not have an eighth-grade math test score, and 17% do not have a ninth-grade score.[5] There are several possible explanations for this outcome: students might have transferred from another district, did not take the exam, or perhaps simply did not have scores appearing in the database. Missing data appear more likely for the subset of students who tend to be male, white or Hispanic, older, and designated as having special education status (and thus potentially exempt from the test). Convincing exclusion restrictions are not available to adequately assess the importance of selection of this type.[6] However, later in the article we show that our quality measure is not correlated with missing test scores, suggesting that this type of selection or gaming of the system does not unduly influence our measure of teacher quality.

Finally, the raw data suggest that racial and income test score gaps rise dramatically between the eighth and ninth grade. While we expect that higher-ability students may gain more in 1 year of education than lower-ability students, we also suspect that the rising gap may be a function of the different exams. In figure 1, kernel density estimates of the eighth- and ninth-grade math test scores are plotted. The ninth-grade scores are skewed right while the eighth-grade test score distribution is more sym-

[5] Eighty-six percent of the students took the TAP (ninth-grade test), and, of this group, we observe scores for 98%.

[6] If selection is based on potential test score improvements because schools and teachers are gaming test score outcomes by reporting scores only for students with the largest gains, we could overstate the impact of teacher quality. Identification of a selection equation requires an exclusion restriction that is able to predict the propensity to have a test score in the administrative records but is not correlated with the educational production function's error term. While there is no obvious candidate, we tried several, including absences, distance to school, and distance to school interacted with whether the student is in their neighborhood school. With the caveat that none of these instruments are ideal, our primary conclusions are unaffected by a selection correction that uses them.

FIG. 1.—Kernel density estimates of eighth- and ninth-grade math test scores. Test scores are measured in grade equivalents. Estimates are calculated using the Epanechnikov kernel. For the eighth-grade test score a bin width of approximately 0.14 is used, while for the ninth-grade test a bin width of approximately 0.26 is used.

metric. As a consequence, controlling for eighth-grade test scores in the regression of ninth-grade test scores on teacher indicators and other student characteristics may not adequately control for the initial quality of a particular teacher's students and may thus lead us to conclude that teachers with better than average students are superior instructors. We drop the top and bottom 1% of the students by change in test scores to partly account for this problem. We also discuss additional strategies, including using alternative test score measures that are immune to differences in scaling of the test, accounting for student attributes, and analyzing groups of students by initial ability.

2. *Classroom Scheduling*

A second important feature of the student data is the detailed scheduling information that allows us to construct the complete history of a student's class schedule while in the CPS high schools. The data include where (room number) and when (semester and period) the class met, the teacher assigned, the title of the class, and the course level (i.e., advanced placement, regular, etc.). Furthermore, we know the letter grade received and the number of classroom absences. Because teachers and students were matched to the same classroom, we have more power to estimate teacher

effects than is commonly available in administrative records where matching occurs at the school or grade level. Additionally, since we have this information for every student, we are able to calculate measures of classroom peers.

One natural concern in estimating teacher quality is whether there are lingering influences from the classroom sorting process. That is, students may be purposely placed with certain instructors based on their learning potential. The most likely scenario involves parental lobbying, which may be correlated with expected test score gains, but a school or teacher may also exert influence that results in nonrandom sorting of students.[7]

To assess the extent to which students may be sorted based on expected test score gains, we calculate test score dispersion for the observed teacher assignments and for several counterfactual teacher assignments. In table 2, we report the degree to which the observed within-teacher standard deviation in students' pre-ninth-grade performance differs from simulated classrooms that are either assigned randomly or sorted based on test score rank. We use three lagged test score measures for assignment: eighth-grade test scores, sixth- to seventh-grade test score gains, and seventh- to eighth-grade test score gains. Each panel reports results for the three fall semesters in our data.[8] The top row of each panel, labeled "Observed," displays the observed average within-teacher standard deviation of these measures. This is the baseline to which we compare the simulations. Each of the four subsequent rows assigns students to teachers either randomly or based on pre-ninth-grade performance.

Row 2 displays the average within-teacher standard deviation when students are perfectly sorted across teachers within their home school.[9] Such a within-school sorting mechanism reduces the within-teacher standard deviation to roughly 20% of the observed analog. In contrast, if we randomly assign students to classrooms within their original school, as shown in row 3, the average within-teacher standard deviation is very close to the within-teacher standard deviation that is observed in the data. Strikingly, there is no evidence that sorting occurs on past gains; the

[7] Informal discussions with a representative of the Chicago public school system suggest that parents have little influence on teacher selection and, conditional on course level, the process is not based on student characteristics. Moreover, our use of first-year high school students may alleviate concern since it is likely more difficult for schools to evaluate new students, particularly on unobservable characteristics.

[8] The estimates for the spring semester are very similar and available from the authors on request.

[9] For example, within an individual school, say there are three classrooms with 15, 20, and 25 students. In the simulation, the top 25 students, based on our pre-ninth-grade measures, would be placed together, the next 20 in the second classroom, and the remainder in the last. The number of schools, teachers, and class sizes are set equal to that observed in the data.

Table 2
Mean Standard Deviation by Teacher of Lagged Student Test Score Measures

	Eighth-Grade Scores (1)	Sixth to Seventh Change (2)	Seventh to Eighth Change (3)
Fall 1997:			
Observed	1.042	.659	.690
Perfect sorting across teachers within school	.214	.132	.136
Randomly assigned teachers within school	1.211	.635	.665
Perfect sorting across teachers	.006	.004	.004
Randomly assigned teachers	1.445	.636	.662
Fall 1998:			
Observed	1.095	.653	.731
Perfect sorting across teachers within school	.252	.151	.175
Randomly assigned teachers within school	1.279	.635	.721
Perfect sorting across teachers	.007	.005	.008
Randomly assigned teachers	1.500	.633	.720
Fall 1999:			
Observed	1.142	.662	.792
Perfect sorting across teachers within school	.274	.168	.217
Randomly assigned teachers within school	1.320	.647	.766
Perfect sorting across teachers	.007	.005	.009
Randomly assigned teachers	1.551	.652	.780

NOTE.—In each cell, we report the average standard deviation by teacher for the lagged math test measure reported at the top of the column when students are assigned to teachers based on the row description. "Observed" calculates the average standard deviation for the observed assignment of students to teachers. "Perfect sorting" assigns students to teachers either within school or across schools based on the test score measure at the top of the column. "Randomly assigned teachers" sort students into teachers either within or across schools based on a randomly generated number from a uniform distribution. The random assignments are repeated 100 times before averaging across all teachers and all random assignments. The top panel reports averages for the fall of 1997, the middle panel for 1998, and the bottom panel for 1999.

observed standard deviations are even slightly larger than the simulations. Using eighth-grade test scores, the randomly assigned matches tend to have within-teacher standard deviations that are roughly 15% higher than the observed assignments. But clearly, the observed teacher dispersion in lagged math scores is much closer to what we would expect with random sorting of students than what we would expect if students were sorted based on their past performance.[10]

Finally, rows 4 and 5 show simulations of perfectly sorted and randomly assigned classrooms across the entire school district. Here, the exercise disregards which school the student actually attends. Consequently, this

[10] These calculations are done using all levels of courses—honors, basic, regular, etc. Because most classes are "regular," the results are very similar when we limit the analysis to regular-level classes.

example highlights the extent to which classroom composition varies across versus within school. We find that the randomly assigned simulation (row 5) is about 18% above the equivalent simulation based solely on within-school assignment and roughly 37% above the observed baseline. Furthermore, there is virtually no variation within randomly assigned classrooms across the whole district. Thus, observed teacher assignment is clearly closer to random than sorted, especially with regard to previous achievement gains, but some residual sorting in levels remains. About half of that is due to within-school classroom assignment and half to across-school variation. School fixed effects provide a simple way to eliminate the latter (Clotfelter, Ladd, and Vigdor 2004).

B. Teacher Records

Finally, we match student administrative records to teacher administrative records using school identifiers and eight-character teacher codes from the student data.[11] The teacher file contains 6,890 teachers in CPS high schools between 1997 and 1999. Although these data do not provide information on courses taught, through the student files we identify 1,132 possible teachers of ninth-grade mathematics classes (these are classes with a "math" course number, although some have course titles suggesting they are computer science). This list is further pared by grouping all teachers who do not have at least 15 student-semesters during our period into a single "other" teacher code for estimation purposes.[12] Ultimately, we identify teacher effects for 783 math instructors, as well as an average effect for those placed in the "other" category. While the student and teacher samples are not as big as those used in some administrative files, they allow for reasonably precise estimation.

Matching student and teacher records allows us to take advantage of a

[11] Additional details about the matching are available in the appendix.

[12] The larger list of teachers incorporates anyone instructing a math class with at least one ninth-grade student over our sample period, including instructors who normally teach another subject or grade. The number of student-semesters for each teacher over 3 years may be smaller than expected for several reasons (this is particularly evident in fig. 2 below). Most obviously, some teacher codes may represent errors in the administrative data. Also, some codes may represent temporary vacancies. More importantly, Chicago Public Schools high school teachers teach students in multiple grades as well as in subjects other than math. In fact, most teachers of math classes in our analysis sample (89%) teach students of multiple grade levels. For the average teacher, 58% of her students are in the ninth grade. In addition, roughly 40% of the teachers in the analysis sample also teach classes that are not math classes. Without excluding students for any reason, the teachers in our sample have an average of 189 unique students in all grades and all subjects. Limiting the classes to math courses drops the average number of students to 169. When we further limit the students to ninth graders, the average number of students is 80.

third feature of the data: the detailed demographic and human capital information supplied in the teacher administrative files. In particular, we can use a teacher's sex, race/ethnicity, experience, tenure, university attended, college major, advanced degree achievement, and teaching certification to decompose total teacher effects into those related to common observable traits of teachers and those that are unobserved, such as drive, passion, and connection with students.

In order to match the teacher data to the student data we have to construct an alphanumeric code in the teacher data similar to the one provided in the student data. The teacher identifier in the student data is a combination of the teacher's position number and letters from the teacher's name, most often the first three letters of his or her last name. We make adjustments to the identifiers in cases for which the teacher codes in the student files do not match our constructed codes in the teacher data due to discrepancies that arise for obvious reasons such as hyphenated last names, use of the first initial plus the first two letters of the last name, or transposed digits in the position number. Ultimately we are unable to resolve all of the mismatches between the student and teacher data but are able to match teacher characteristics to 75% of the teacher codes for which we estimate teacher quality (589 teachers). Table 3 provides descriptive statistics for the teachers we can match to the student administrative records. The average teacher is 45 years old and has been in the CPS for 13 years. Minority math and computer science teachers are underrepresented relative to the student population, as 36% are African American and 10% Hispanic, but they compare more favorably to the overall population of Chicago, which is 37% black or African American and 26% Hispanic or Latino (Census 2000 Fact Sheet for Chicago, U.S. Census Bureau). Eighty-two percent are certified to teach high school, 37% are certified to be a substitute, and 10%–12% are certified to teach bilingual, elementary, or special education classes. The majority of math teachers have a master's degree, and many report a major in mathematics (48%) or education (18%).[13]

III. Basic Empirical Strategy

In the standard education production function, achievement, Y, of student i with teacher j in school k at time t is expressed as a function of cumulative own, family, and peer inputs, X, from age 0 to the current

[13] Nationally, 55% of high school teachers have a master's degree, 66% have an academic degree (e.g., mathematics major), and 29% have a subject area education degree (U.S. Department of Education 2000).

Table 3
Descriptive Statistics for the Teachers Matched to Math Teachers in the Student Data

	Mean	Standard Deviation
Demographics:		
Age	45.15	10.54
Female	.518	.500
African American	.360	.480
White	.469	.499
Hispanic	.100	.300
Asian	.063	.243
Native American	.007	.082
Human capital:		
BA major: education	.182	.386
BA major: all else	.261	.440
BA major: math	.484	.500
BA major: science	.073	.260
BA university, *US News* 1	.092	.289
BA university, *US News* 2	.081	.274
BA university, *US News* 3	.151	.358
BA university, *US News* 4	.076	.266
BA university, *US News* 5	.019	.135
BA university, *US News* else	.560	.497
BA university missing	.020	.141
BA university local	.587	.493
Master's degree	.521	.500
PhD	.015	.123
Certificate, bilingual education	.119	.324
Certificate, child	.015	.123
Certificate, elementary	.100	.300
Certificate, high school	.823	.382
Certificate, special education	.107	.309
Certificate, substitute	.365	.482
Potential experience	19.12	11.30
Tenure at CPS	13.31	10.00
Tenure in position	5.96	6.11
Number of observations	589	

NOTE.—There are 783 teachers identified from the student estimation sample that have at least 15 student-semesters for math classes over the 1997–99 sample period. The descriptive statistics above apply to the subset of these teachers that can be matched to the teacher administrative records from the Chicago Public Schools. *US News* rankings are from U.S. News & World Report (1995): level 1 = top tier universities (top 25 national universities + tier 1 national universities) + (top 25 national liberal arts colleges + tier 1 national liberal arts colleges); level 2 = second tier national universities + second tier national liberal arts colleges; level 3 = third tier national universities + third tier national liberal arts colleges; level 4 = fourth tier national universities + fourth tier national liberal arts colleges; and level 5 = top regional colleges and universities.

age, as well as cumulative teacher and school inputs, S, from grades kindergarten through the current grade:

$$Y_{ijkt} = \beta \sum_{t=-5}^{T} X_{it} + \gamma \sum_{t=0}^{T} S_{ijkt} + \varepsilon_{ijkt}. \tag{1}$$

The requirements to estimate (1) are substantial. Without a complete set of conditioning variables for X and S, omitted variables may bias estimates of the coefficients on observable inputs unless strong and unlikely as-

sumptions about the covariance structure of observables and unobservables are maintained. Thus, alternative identification strategies are typically applied.

A simple approach is to take advantage of multiple test scores. In particular, we estimate a general form of the value-added model by including eighth-grade test scores as a covariate in explaining ninth-grade test scores. Lagged test scores account for the cumulative inputs of prior years while allowing for a flexible autoregressive relationship in test scores. Controlling for past test scores is especially important with these data, as information on the family and pre-ninth-grade schooling is sparse.

We estimate an education production model of the general form

$$Y_{ikt}^{9} = \alpha Y_{it-1}^{8} + \beta X_i + \tau T_{it} + \theta_i + \rho_k + \varepsilon_{ijkt}, \tag{2}$$

where Y_{ikt}^{9} refers to the ninth-grade test score of student i, who is enrolled in ninth grade at school k in year t; Y_{it-1}^{8} is the eighth-grade test score for student i, who is enrolled in ninth grade in year t; and θ_i, ρ_k, and ε_{ijk} measure the unobserved impact of individuals, schools, and white noise, respectively.[14] Each element of matrix T_{it} records the number of semesters spent in a math course with teacher j. To be clear, this is a cross-sectional regression estimated using ordinary least squares with a slight deviation from the standard teacher fixed effect specification.[15] Therefore, τ_j is the jth element of the vector τ, representing the effect of one semester spent with math teacher j. Relative to equation (1), the impact of lagged schooling and other characteristics is now captured by the lagged test score measure.

While the value-added specification helps control for the fact that teachers may be assigned students with different initial ability on average, this strategy may still mismeasure teacher quality. For simplicity, assume that all students have only one teacher for one semester so that the number of student-semesters for teacher j equals the number of students, N_j. In this case, estimates of τ_j may be biased by $\rho_k + \frac{1}{N_j}\sum_{i=1}^{N_j}\theta_i + \frac{1}{N_j}\sum_{i=1}^{N_j}\varepsilon_{ijkt}$.

The school term ρ_k is typically removed by including measures of school quality, a general form of which is school fixed effects. School fixed effects estimation is useful to control for time-invariant school characteristics that covary with individual teacher quality, without having to attribute the school's contribution to specific measures. However, this strategy requires the identification of teacher effects to be based on differences in the number of semesters spent with a particular teacher and teachers that switch schools during our 3-year period. For short time periods, such as

[14] All regressions include year indicators to control for any secular changes in test performance or reporting.

[15] For repeaters, we use their first ninth-grade year so as to allow only a 1-year gap between eighth- and ninth-grade test results.

a single year, there may be little identifying variation to work with. Thus, this cleaner measure of the contribution of mathematics teachers comes at the cost of potential identifying variation. In addition, to the extent that a principal is good because she is able to identify and hire high quality teachers, some of the teacher quality may be attributed to the school. For these reasons, we show many results both with and without allowing for school fixed effects.

Factors influencing test scores are often attributed to a student's family background. In the context of gains, many researchers argue that time-invariant qualities are differenced out, leaving only time-varying influences, such as parental divorce or a student's introduction to drugs, in $\frac{1}{N_j}\sum_{i=1}^{N_j}\theta_i$. While undoubtedly working in gains lessens the omitted variables problem, we want to be careful not to claim that value-added frameworks completely eliminate it. In fact, it is quite reasonable to conjecture that student gains vary with time-varying unobservables. But given our statistical model, bias is only introduced to the teacher quality rankings if students are assigned to teachers based on these unobservable changes.[16] Furthermore, we include a substantial list of observable student, family, and peer traits because they may be correlated with behavioral changes that influence achievement and may account for group differences in gain trajectories.

Finally, as the findings of Kane and Staiger (2002) make clear, the error term $\frac{1}{N_j}\sum_{i=1}^{N_j}\varepsilon_{ijkt}$ is particularly problematic when teacher fixed effect estimates are based on small populations (small N_j). In this case, sampling variation can overwhelm signal, causing a few good or bad draws to strongly influence the estimated teacher fixed effect. Consequently, the standard deviation of the distribution of estimated τ_j is most likely inflated.

This problem is illustrated by figure 2, in which we plot our estimates $\hat{\tau}_j$ (conditional on eighth-grade math score, year indicators, and student, family, and peer attributes, as described below) against the number of student-semesters on which the estimate is based. What is notable is that the lowest and highest performing teachers are those with the fewest student-semesters. The expression $\sum_i T_{ij}$ represents the number of student-semesters taught by teacher j over the 3-year period examined (see n. 12 for a discussion of the distribution of $\sum_i T_{ij}$). As more student-semesters are used to estimate the fixed effect, the importance of sampling variation declines and reliability improves. Regressing $|\hat{\tau}_j|$ on $\sum_i T_{ij}$ summarizes this association. Such an exercise has a coefficient estimate of -0.00045 with a standard error of 0.000076, suggesting that number of student-semesters is a critical correlate of the magnitude of estimated teacher quality. The

[16] We do not discount the possibility of this type of sorting, especially for transition schools, which are available to students close to expulsion. School fixed effects pick this up, but we also estimate the results excluding these schools.

FIG. 2.—Teacher effect estimates versus student counts

association declines as we raise the minimum threshold on $\sum_i T_{ij}$ and completely disappears when $\sum_i T_{ij} \geq 250$.[17]

To address the problem of sampling error, we analytically adjust the variance of $\hat{\tau}_j$ for the size of the sampling error by assuming that the estimated teacher fixed effect is the sum of the true teacher effect, τ_j, plus some error, ε_j, where ε_j is uncorrelated with τ_j. While we would like to estimate σ_τ^2, the variance of the estimated teacher effects is $\sigma_{\hat{\tau}}^2 = \sigma_\tau^2 + N^{-1}\varepsilon\varepsilon$. That is, the variance of the estimated teacher effects has two components—the true variance of the teacher effects and average sampling variance. We use the mean of the square of the standard error estimates of $\hat{\tau}_j$ as an estimate of the sampling error variance and subtract this from the observed variance of $\hat{\tau}_j$ to get an adjusted variance, σ_τ^2. We report the associated standard deviations, σ_τ and $\sigma_{\hat{\tau}}$, in subsequent tables. We also show how these values vary as we increase the minimum evaluation threshold, $\sum_i T_{ij}$. For statistical models that include school fixed effects, we estimate that roughly 30% of the standard deviation in estimated teacher quality is due to sampling error. If we raise the minimum number

[17] When $\sum_i T_{ij} \geq 250$, the point estimate and standard error are −0.0000367 (−0.0001597). While the standard error doubles due to the smaller sample of teachers as we move from the student threshold from 15 to 250, the point estimate declines substantially as well.

of student-semesters to identify an individual teacher to 200, only 14% of the standard deviation in teacher quality is due to sampling error.[18]

In the section to follow, we present our baseline estimates that ignore the existence of most of these potential biases. We then report results that attempt to deal with each potential bias. To the extent that real world evaluation might not account for these problems, this exercise could be considered a cautionary tale of the extent to which teacher quality estimates can be interpreted incorrectly.

Finally, we examine whether teacher quality can be explained by demographic and human capital attributes of teachers. Because of concerns raised by Moulton (1986) about the efficiency of ordinary least squares (OLS) estimates in the presence of school-specific fixed effects and because students are assigned multiple teachers per year, we do not include the teacher characteristics directly in equation (2). Rather, we employ a generalized least squares (GLS) estimator outlined in Borjas (1987) and Borjas and Sueyoshi (1994). This estimator regresses $\hat{\tau}_j$ on teacher characteristics Z:

$$\hat{\tau}_j = \phi Z_j + u_j. \tag{3}$$

The variance of the errors is calculated as the covariance matrix derived from OLS estimates of (3) and the portion of equation (2)'s variance matrix related to the $\hat{\tau}$ coefficient estimates, V.

$$\Omega = \sigma_u^2 I_J + V. \tag{4}$$

The Ω term in (4) is used to compute GLS estimates of the observable teacher effects.

IV. Results

A. The Distribution of Teacher Quality

Our naive baseline estimates of teacher quality are presented in table 4. In column 1 we present details on the distribution of $\hat{\tau}_j$, specifically the standard deviation and the 10th, 25th, 50th, 75th, and 90th percentiles. We also list the p-value for an F-test of the joint significance of the teacher effects (i.e., $\hat{\tau}_j = 0$ for all j) and the p-value for an F-test of the other regressors. In this parsimonious specification, the list of regressors is limited to eighth-grade math scores, year dummies, and indicators of the test

[18] Note, however, that excluding teachers with small numbers of students is limiting because new teachers, particularly those for whom tenure decisions are being considered, may not be examined. This would be particularly troubling for elementary school teachers with fewer students per year.

Table 4
Distribution of the Estimated Teacher Effects

	Distribution of Teacher Fixed Effects	
	Unweighted (1)	Weighted (2)
10th percentile	−.38	−.33
25th percentile	−.24	−.19
50th percentile	−.08	−.05
75th percentile	.17	.17
90th percentile	.53	.53
90–10 gap	.91	.86
75–25 gap	.41	.36
Standard deviation	.398	.354
Adjusted standard deviation	.355	
Adjusted R^2	.69	
p-value for the F-test on:		
Teacher fixed effects	.000	
Eighth-grade math score and year dummies	.000	
Math scores units	Grade equivalents	
Number of students	52,957	
Number of teachers	783	
Number of students threshold	15	

NOTE.—All results are based on a regression of ninth-grade math test score on eighth-grade math test score, ninth-grade test score level, eighth-grade test score level, an indicator equal to one if the information on eighth-grade test score level is missing, teacher semester counts, and year indicators.

level and format.[19] Clearly, we cannot rule out the importance of confounding changes in family, student, peer, and school influences as well as random fluctuations in student performance across teachers. Rather,

[19] Naturally, the key covariate in our production functions, regardless of specification, is the eighth-grade test score. The t-statistic on this variable often exceeds 200. Yet the magnitude of the point estimate is somewhat surprising in that it is often greater than 1. For example, in our sparsest specification, the coefficient on eighth-grade test score is 1.30 (0.01). This suggests that the math test score time series may not be stationary. However, this is not likely to be a problem since we are working off of the cross-section. It would become an issue if we were to include longitudinal information on tenth or eleventh grade. Nevertheless, a simple way to deal with nonstationarity is to estimate eq. (2) in differenced form. Such a specification will lead to inconsistent estimates because of the correlation between the error term and the lagged differenced dependent variable, but a common strategy to avoid this problem is to use the twice lagged differenced dependent variable, in our case the difference between seventh- and sixth-grade scores, as an instrument. This instrumental variables estimator reduces our estimates of the dispersion in teacher effects slightly (by less than 0.02 in our preferred specifications) but broadly supports the results presented below. It also suggests that controlling for student fixed effects is not likely to change our results significantly. However, we do not want to stress this result too much since it is based on a potentially nonrepresentative sample, those with test scores in every year between sixth and ninth grade.

Table 5
Quartile Rankings of Estimated Teacher Effects in Years t and $t + 1$: Percent of Teachers by Row

	Quartile in Year $t + 1$			
Quartile in year t:	1	2	3	4
1	36	29	26	10
2	24	31	32	12
3	20	32	23	24
4	8	12	23	57

NOTE.—χ^2 test of random quartile assignment: $p < .000$. Quartile rankings are based on teacher effects estimated for each year based on the specification in col. 1 of table 6.

we report these estimates as a baseline for considering the importance of these biases.

Consequently, the estimated range of the teacher fixed effects is quite broad, perhaps implausibly so. The standard deviation of $\hat{\tau}$ is 0.40 with gaps between the 90th percentile and 10th percentile teacher of 0.9 grade equivalents. Furthermore, approximately 0.4 grade equivalents separate average gains between the 75th and 25th percentile teacher. An F-test of the joint significance of $\hat{\tau}$ easily rejects no teacher effects at the highest significance level.

Because we have multiple classrooms per teacher and can follow teachers across years, the robustness of these results can be explored by tracking the stability of individual teacher quality over time. To do so, we simply estimate equation (2) separately by school year and then compare estimates for the same teacher in different school years. The extent to which $\hat{\tau}_{jt}$ is autocorrelated gives a measure of the extent to which signal dominates noise in our quality rankings.

Table 5 displays one such analysis. Here we report a transition matrix linking quartile rankings of $\hat{\tau}_{jt}$ with quartile rankings of $\hat{\tau}_{jt+1}$. Quartile 1 represents the lowest 25% of teachers as ranked by the teacher quality estimate, and quartile 4 represents the highest 25%. The table reports each cell's share of the row's total or the fraction of teachers in quartile q in year t that move to each of the four quartiles in year $t + 1$. If our estimates are consistent with some signal, whether it is quality or something correlated with quality, we would expect masses of teachers on the diagonals of the transition matrix. We expect cells farther from the diagonals to be monotonically less common. Particularly noisy estimates would not be able to reject the pure random assignment result that each cell would contain equal shares of teachers. In this rather extreme case, teachers would be randomly assigned a new quality ranking each year, and the correlation between this year's ranking and the next would be 0.

Our results suggest a nontransitory component to the teacher quality measure. Of the teachers in the lowest quality quartile in year t, 36%

remain in year $t + 1$, 29% move into quartile 2, 26% into quartile 3, and 10% into the highest quartile. Of those in the highest quartile in year t (row 4), 57% remain the following year, 23% move one category down, and only 20% fall into the lowest half of the quality distribution. A chi-square test easily rejects random assignment.[20]

Moreover, we have also explored to what extent teachers in the top and bottom deciles of the quality distribution continue to rank there the following year. Of the teachers in the top decile, 56% rank there the following year. This is highly significant relative to the random draw scenario whereby 10% would again appear in the top decile in consecutive years. However, of those teachers in the bottom decile, only 6% remain there the following year. Given our sample sizes, this is not significantly different from the random assignment baseline.

We believe the latter result is partly driven by greater turnover among teachers in the bottom decile. To appear in our transition matrix, a teacher must be in the administrative records for two consecutive years. Therefore, if poor performing teachers are more likely to leave the school system, our test is biased; the random draw baseline would no longer be 10%. To investigate this possibility, we regress an indicator of whether the teacher appears in the teacher records in year $t + 1$ on whether she is ranked in the top or bottom decile of the quality distribution in year t.[21] We find that a teacher ranked at the bottom is 13% less likely (standard error of 6%) than a teacher ranked in the 10th to 90th percentile to appear in the administrative records the following year. In contrast, teacher turnover for those in the top decile is no different than turnover for the 10th to 90th percentile group. While accounting for the higher turnover rate of bottom decile ranked teachers does not lead us to conclude that there is significant persistence at the very bottom of the quality distribution in this particular specification, it does once we begin to refine the production function specification below.

Regardless, all of these results emphasize that teacher quality evaluated using parsimonious specifications with little attention to measurement issues still has an important persistent component. However, the transitory part, which is aggravated by sampling error when looking at estimates based on one year, is also apparent. Furthermore, the magnitude of the estimates is perhaps improbably large.

[20] Similarly, regressing contemporaneous teacher quality on lagged teacher quality results in a point estimate of 0.47 (0.04) for 1998 and 0.62 (0.07) for 1999. Limiting it to teachers in all 3 years, the coefficients (and standard errors) on lagged and twice lagged teacher quality are 0.49 (0.10) and 0.25 (0.09).

[21] Unfortunately, we cannot distinguish quits from layoffs or exits out of teaching from exits into other school systems.

B. The Impact of Sampling Error

We next consider how sampling error may affect our results. We already attempt to improve the signal-to-noise ratio by throwing out students with test score changes in the extreme tails and by restricting identified teachers to those with more than 15 student-semesters. However, Kane and Staiger (2002) show that more than one-half of the variance in score gains from small North Carolina schools (typically smaller than our counts of student-semesters, $\sum_i T_{ij}$) and one-third of the variance in test score gains from larger North Carolina schools are due to sampling variation. Figure 2 emphasizes the susceptibility of our results to these concerns as well.

The row labeled "Adjusted Standard Deviation" in table 4 presents an estimate of σ_τ, the true standard deviation of the teacher effects after adjusting for sampling variation as described earlier. This modification reduces the standard deviation from 0.40 to 0.36. We can confirm this result simply by adjusting for possible overweighting of unreliable observations. Column 2 reports the distribution of $\hat{\tau}_j$, when weighted by $\sum_i T_{ij}$. The weighted standard deviation of the teacher effects drops to 0.35, virtually identical to the adjusted standard deviation reported in column 1. In either case, we conclude that dispersion in teacher quality is wide and educationally significant.

C. Family, Student, and Peer Characteristics

The teacher quality results reported thus far are based on parsimonious specifications. They do not fully capture heterogeneity in student, family, and peer background that could be correlated with particular teachers. In table 6 we report results in which available student, family, and peer group characteristics are included. For comparison purposes, column 1 repeats the findings from table 4. In each column we report unadjusted, adjusted, and weighted standard deviation estimates, as well as p-values for F-tests of the joint significance of the teacher effects and the other regressors as they are added to the production function.

In column 2 we incorporate student characteristics including sex, race, age, designated guardian relationship (mom, dad, stepparent, other relative, or nonrelative), and free and reduced-price lunch eligibility. In addition, we include a measure of the student's average ninth-grade math class size, as is standard in educational production analysis, and controls for whether the student changed high school or repeated ninth grade.[22]

[22] Jointly these background measures are quite significant; individually, the sex and race measures are the primary reason. The ninth-grade scores for female students are 0.16 (0.01) less than males, and African American and Hispanic students score 0.50 (0.03) and 0.31 (0.03) less than non–African American, non-Hispanic students. Accounting for additional student characteristics such as dis-

Table 6
Distribution of the Estimated Teacher Effects

	(1)	(2)	(3)	(4)	(5)
Standard deviation	.398	.384	.298	.303	.273
Adjusted standard deviation	.355	.341	.242	.230	.193
Weighted standard deviation	.354	.335	.246	.248	.213
p-value, *F*-test of teacher effects	.000	.000	.000	.000	.000
p-value, *F*-test of lagged test score and year	.000				
p-value, *F*-test for basic student covariates		.000			
p-value, *F*-test for school effects				.000	.000
p-value, *F*-test for additional student, peer, and neighborhood covariates			.000		.000
Included covariates:					
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Basic student covariates	No	Yes	Yes	Yes	Yes
Additional student covariates	No	No	Yes	No	Yes
Math peer covariates	No	No	Yes	No	Yes
Neighborhood covariates	No	No	Yes	No	Yes
School fixed effects	No	No	No	Yes	Yes
Number of students threshold	15	15	15	15	15

NOTE.—All results are based on a regression of ninth-grade math test score on eighth-grade math test score, teacher student-semester counts, year indicators, ninth-grade test level, eighth-grade test level, an indicator equal to one if the information on eighth-grade test score level is missing, and other covariates as listed in the table. All test scores are measured in grade equivalents. Basic student covariates include gender, race, age, guardianship, number of times in ninth grade, free or reduced-price lunch status, whether changed school during school year, and average math class size. Additional student covariates include level and subject of math classes, cumulative GPA, class rank, disability status, and whether school is outside of the student's residential neighborhood. Peer covariates include the 10th, 50th, and 90th percentile of math class absences and eighth-grade math test scores in ninth-grade math classes. Neighborhood covariates include median family income, median house value, and fraction of adult population that fall into five education categories. All neighborhood measures are based on 1990 census tract data. There are 52,957 students and 783 teachers in each specification.

These controls reduce the size of the adjusted standard deviation by a small amount, but the estimates remain large and highly statistically significant.

In column 3 we introduce additional student controls, primarily related to performance, school choice, peers, and neighborhood characteristics. The additional student regressors are the level and subject matter of math classes, the student's cumulative grade point average, class rank, and disability status, and whether the school is outside of her residential neigh-

ability status and average grades, neighborhood characteristics, and peer controls reduces the racial gaps markedly, but the female gap nearly doubles. Students whose designated guardian is the father have, on average, 0.10–0.20 higher test scores than do students with other guardians, but these gaps decline substantially with other controls. Math class size has a positive and significant relationship with test scores that becomes negative and statistically significant once we include the col. 3 controls.

borhood.[23] The neighborhood measures are based on Census data for a student's residential census tract and include median family income, median house value, and the fraction of adults that fall into five education categories. These controls are meant to proxy for unobserved parental influences. Again, like many of the student controls, the value-added framework should, for example, account for permanent income gaps but not for differences in student growth rates by parental income or education. Finally, the math class peer characteristics are the 10th, 50th, and 90th percentiles of absences, as a measure of disruption in the classroom, and the same percentiles of eighth-grade math test scores, as a measure of peer ability. Because teacher ability may influence classroom attendance patterns, peer absences could confound our estimates of interest, leading to downward biased teacher quality estimates.[24]

Adding student, peer, and neighborhood covariates reduces the adjusted standard deviation to 0.24, roughly two-thirds the size of the naive estimates reported in column 1.[25] Much of the attenuation comes from adding either own or peer performance measures. Nevertheless, regardless of the controls introduced, the dispersion in teacher quality remains large and statistically significant.

Once again, transition matrices for the full control specification clearly reject random quality draws. The quartile-transition matrix is reported in

[23] We also experiment with additional controls for student ability, including eighth-grade reading scores, sixth- and seventh-grade math scores, higher-order terms (square and cube) and splines in the eighth-grade math score, and the variance in sixth to eighth-grade math scores. Compared to the col. 3 baseline, the largest impact is from adding the higher-order terms in eighth-grade scores. This reduces the adjusted standard deviation by just under 0.03. When school fixed effects are also included, the largest impact of any of these specification adjustments is half that size.

[24] See Manski (1993) for a methodological discussion and Hoxby (2000) and Sacerdote (2001) for evidence. While we hesitate to place a causal interpretation on the peer measures, there is a statistical association between a student's performance and that of her peers. The point estimates (standard errors) on the 10th, 50th, and 90th percentile of peer absences are 0.009 (0.005), −0.002 (0.002), and −0.002 (0.0007). Thus it appears that the main statistically significant association between own performance and peer absences is from the most absent of students. The point estimates on the 10th, 50th, and 90th percentile of eighth-grade math scores are 0.028 (0.013), 0.140 (0.025), and 0.125 (0.019). These peer measures reduce the student's own eighth-grade math test score influence by 17% and suggest that high performers are most associated with a student's own performance.

[25] Arguably, part of the reduction in variance is excessive, as teachers may affect academic performance through an effect on absences or GPA. About half of the reduction in teacher dispersion between cols. 2 and 3 (adding peer and own student performance and schooling measures) is due to peer measures. That said, when we identify teacher effects within-school, peer measures have little additional power in explaining teacher quality dispersion.

Table 7
Quartile Rankings of Estimated Teacher Effects in Years *t* and *t* + 1: Percent of Teachers by Row

	Quartile in Year $t + 1$			
Quartile in year t:	1	2	3	4
1	33	32	16	19
2	32	25	31	13
3	17	25	33	26
4	15	21	23	41

NOTE.—χ^2 test of random quartile assignment: $p < .001$. Quartile rankings are based on teacher effects estimated for each year based on the specification including lagged math test score, year indicators, and all student, peer, and neighborhood covariates (col. 3 of table 6).

table 7. Forty-one percent of teachers ranking in the top 25% in one year rank in the top 25% in the following year. Another 23% slip down one category, 21% two categories, and 15% to the bottom category.[26]

D. Within-School Estimates

Within-school variation in teacher quality is often preferred to the between-school variety as it potentially eliminates time-invariant school-level factors. In our case, since we are looking over a relatively short window (3 years), this might include the principal, curriculum, school size or composition, quality of other teachers in the school, and latent family or neighborhood-level characteristics that can influence school choice. Because our results are based on achievement gains, we are generally concerned only with changes in these factors. Therefore, restricting the source of teacher variation to within-school differences will result in a more consistent, but less precisely estimated, measure of the contribution of teachers.

Our primary method of controlling for school-level influences is school fixed effects. As mentioned above, identification depends on differences in the intensity of students' exposure to different teachers within schools, as well as teachers switching schools during the sample period.[27] We report these results in columns 4 and 5 of table 6. Relative to the analogous columns without school fixed effects, the dispersion in teacher quality and precision of the estimates decline. For example, with the full set of student controls, the adjusted standard deviation drops from 0.24 (col. 3)

[26] Twenty-six percent and 19% of those in the top and bottom deciles remain the next year. Nineteen percent and 14% rated in the top and bottom deciles in 1997 are still there in 1999. Again, turnover is 15% higher among the lowest performing teachers. Adjusting for this extra turnover , the p-value on the bottom decile transition F-test drops from 0.14 to 0.06.

[27] Of the teachers with at least 15 student-semester observations, 69% appear in one school over the 3 years and 18% appear in two schools. Additionally, 13%–17% of teachers in each year show up in multiple schools.

Table 8
Correlation between Teacher Quality Estimates across Specifications

	Minimum Number of Student-Semesters Required to Identify a Teacher		
Specification Relative to Baseline	15 (1)	100 (2)	200 (3)
(0) Baseline	1.00	1.00	1.00
(1) Drop neighborhood covariates	1.00	1.00	1.00
(2) Drop peer covariates	.97	.98	.99
(3) Drop additional student covariates	.92	.93	.94
(4) Drop basic student covariates	.99	1.00	1.00
(5) Drop basic and additional student, peer, and neighborhood characteristics	.88	.85	.87
(6) Drop school fixed effects	.86	.68	.65
(7) Drop school fixed effects and basic and additional student, peer, and neighborhood characteristics	.62	.44	.45
Number of teachers	783	317	122

NOTE.—The col. 1 baseline corresponds to the results presented in col. 5 of table 6. Columns 2 and 3 correspond to the results presented in table 9, cols. 2 and 3, respectively. All specifications include the eighth-grade math test score, teacher student-semester counts, year indicators, the ninth-grade test level, the eighth-grade test level, an indicator equal to one if the information on eighth-grade test schore level is missing, and a constant. The baseline specification additionally includes basic and additional student characteristics, neighborhood and peer characteristics, and school fixed effects. All other specifications include a subset of these controls as noted in the table. See table 6 for the specific variables in each group.

to 0.19 (col. 5), roughly one-half the impact from the unadjusted value-added model reported in column 1. Again, an F-test rejects that the within-school teacher quality estimates jointly equal zero at the 1% level. We have also estimated column 4 and 5 models when allowing for school-specific time effects, to account for changes in principals, curricula, and other policies, and found nearly identical results. The adjusted standard deviations are 0.23 and 0.18, respectively, just 0.01 lower than estimates reported in the table.

Notably, however, once we look within schools, sampling variation accounts for roughly one-third of the unadjusted standard deviation in teacher quality. Furthermore, sampling variation becomes even more problematic when we estimate year-to-year transitions in quality, as in tables 5 and 7, with specifications that control for school fixed effects.

E. Robustness of Teacher Quality Estimates across Specifications

One critique of using test score based measurements to assess teacher effectiveness has been that quality rankings can be sensitive to how they are calculated. We suspect that using measures of teacher effectiveness that differ substantially under alternative, but reasonable, specifications of equation (2) will weaken program incentives to increase teacher effort in order to improve student achievement. To gauge how sensitive our results are to the inclusion of various controls, table 8 reports the robustness of

the teacher rankings to various permutations of our baseline results (col. 5 of table 6). In particular, we present the correlations of our teacher quality estimate based on our preferred statistical model—which controls for school fixed effects as well as student, peer, and neighborhood characteristics—with estimates from alternative specifications.

Because the estimation error is likely to be highly correlated across specifications, we calculate the correlation between estimates using empirical Bayes estimates of teacher effects (e.g., Kane and Staiger 2002). We rescale the OLS estimates using estimates of the true variation in teacher effects and the estimation error as follows:

$$\tau_j^* = \hat{\tau}_j \cdot \frac{\sigma_\tau^2}{\sigma_\tau^2 + \hat{\sigma}_\varepsilon^2}, \tag{5}$$

where $\hat{\tau}_j$ is our OLS estimate of the value added by teacher j, σ_τ^2 is our estimate of the true variation in teacher effects (calculated as described above), and $\hat{\sigma}_\varepsilon^2$ is the noise associated with the estimate of teacher j's effect, namely, the estimation error for $\hat{\tau}_j$. To further minimize concern about sampling variability, we also look at correlations across specifications estimated from samples of teachers that have at least 100 or 200 students during our period.

In rows 1–4, we begin by excluding, in order, residential neighborhood, peer, student background, and student performance covariates. Individually, each of these groups of variables has little impact on the rankings. Teacher rankings are always correlated at least 0.92 with the baseline. Even when we drop all of the right-hand-side covariates, except school fixed effects, row 5 shows that the teacher ranking correlations are still quite high, ranging from 0.85 to 0.88.

Only when school fixed effects are excluded is there a notable drop in the correlation with the baseline. In row 6, we exclude school fixed effects but leave the other covariates in place. The teacher quality correlation falls to between 0.65 and 0.86. Excluding the other right-hand-side covariates causes the correlation to fall to between 0.44 and 0.62. That is, without controlling for school fixed effects, rankings become quite sensitive to the statistical model. But as long as we consider within-school teacher quality rankings using a value-added specification, the estimates are highly correlated across specifications, regardless of the other controls included.

Importantly, our results imply that teacher rankings based on test score gains are quite robust to the modeling choices that are required for an individual principal to rank her own teachers. But a principal may have more difficulty evaluating teachers outside her home school. More generally, value-added estimates that do not account for differences across

schools may vary widely based on specification choices which in turn may weaken teacher performance incentives.

F. Additional Robustness Checks

Thus far, we have found that teacher quality varies substantially across teachers, even within the same school, and is fairly robust across reasonable value-added regression specifications. This section provides additional evidence on the robustness of our results to strategic test score reporting, sampling variability, test normalization, and the inclusion of other teachers in the math score production function.

1. *Cream Skimming*

One concern is that teachers or schools discourage some students from taking exams because they are expected to perform poorly. If such cream skimming is taking place, we might expect to see a positive correlation between our teacher quality measures τ_j and the share of teacher j's students that are missing ninth-grade test scores. In fact, we find that this correlation is small (−0.02), opposite in sign to this cream-skimming prediction, and not statistically different from zero.

Another way to game exam results is for teachers or schools to test students whose scores are not required to be reported and then report scores only for those students who do well. To examine this possibility, we calculate the correlation between teacher quality and the share of students excluded from exam reporting.[28] In this case, evidence is consistent with gaming of scores; the correlation is positive (0.07) and statistically different from zero at the 6% level. To gauge the importance of this finding for our results, we reran our statistical models, dropping all students for whom test scores may be excluded from school and district reporting. This exclusion affected 6,361 students (12% of the full sample) but had no substantive impact on our results.

2. *Sampling Variability: Restrictions on Student-Semester Observations*

A simple strategy for minimizing sampling variability is to restrict evaluation to teachers with a large number of student-semesters. In table 9, we explore limiting assessment of teacher dispersion to teachers with at least 50, 100, or 200 student-semesters. We emphasize that a sampling restriction, while useful for its simplicity, can be costly in terms of inference. Obviously, the number of teachers for whom we can estimate quality is reduced. There may also be an issue about how representative

[28] The student test file includes an indicator for whether the student's test score may be excluded from reported school or citywide test score statistics because, e.g., the student is a special or bilingual education student.

Table 9
Further Evidence on the Distribution of the Estimated Teacher Effects

	Student Threshold			Test Scores Measured in Percentiles	Trimming Top and Bottom 3% in Changes
	50 (1)	100 (2)	200 (3)	(4)	(5)
Dependent variable mean	9.21	9.21	9.21	37.88	9.08
Mean test score gain	1.14	1.14	1.14	−2.08	1.06
Number of teachers	508	317	122	783	773
Number of students	52,957	52,957	52,957	52,957	50,392
Without school effects:					
Standard deviation of teacher effects	.233	.227	.193	2.66	.262
Adjusted standard deviation	.205	.211	.180	2.06	.203
Weighted standard deviation	.223	.216	.188	2.22	.211
p-value, *F*-test for teacher effects	.000	.000	.000	.000	.000
With school effects:					
Standard deviation of teacher effects	.192	.183	.154	2.57	.244
Adjusted standard deviation	.143	.155	.133	1.75	.161
Weighted standard deviation	.182	.176	.152	2.04	.188
p-value, *F*-test for teacher effects	.000	.000	.000	.000	.000

NOTE.—See notes to table 6. All regressions include the student, peer, and neighborhood covariates included in the table 6, cols. 3 and 5, specifications.

the teachers are, particularly since we overlook an important sample of teachers—new instructors with upcoming tenure decisions—in addition to teachers who teach multiple grades or nonmath classes. Finally, sampling variation exists with large numbers of students as well, so we would not expect to completely offset concerns about sampling error by simply setting a high minimum count of student-semesters.

Panel A of table 9 includes all covariates from the specification presented in column 3 of table 6. Panel B adds school fixed effects (i.e., col. 5 of table 6). Using a 50, 100, or 200 student-semester threshold, we find that the adjusted standard deviation is roughly 0.18–0.21 without school fixed effects and 0.13–0.15 grade equivalents with school fixed effects. In both cases, the teacher effects are jointly statistically significant. Note that increasing the minimum student-semesters from 15 to 200 increases the average number of student-semesters per teacher from 109 to 284. Consequently, sampling variability drops substantially, from an adjustment of 0.081 (0.273 − 0.192) for the 15-student threshold to 0.021 (0.155 − 0.134) for the 200-student threshold.

3. *More on Test Score Normalization and the Undue Influence of Outliers*

The remaining columns of table 9 include attempts to minimize the influence of outlier observations. Column 4 reports findings using national percentile rankings that are impervious to the normalization problem inherent in grade-equivalent scores.[29] We find that the adjusted standard deviation of $\hat{\tau}_j$ is 1.75 percentile points, a result that is statistically and educationally significant and broadly consistent with the grade-equivalent results.

In the next column, we simply trim the top and bottom 3% of the distribution of eighth- to ninth-grade math test gains from the student sample. We would clearly expect that this sample restriction would reduce the variance, as it eliminates roughly 2,600 students in the tails of the score distribution. Still, the adjusted teacher standard deviation remains large in magnitude and statistically significant at 0.16 grade equivalents.[30]

4. *Including Other Teachers in the Production Function*

We explore one final specification that takes advantage of the detailed classroom scheduling in our data by including a full set of English teacher semester counts, akin to the math teacher semester count, T_i, in equation (2). Assuming that the classroom-sorting mechanism is similar across subject areas (e.g., parents who demand the best math teacher will also demand the best English teacher or schools will sort students into classrooms and assign classes to teachers based on the students' expected test score gains), the English teachers will pick up some sorting that may confound estimates of τ. Moreover, the English teachers may help us gauge the importance of teacher externalities, that is, the proverbial superstar teacher who inspires students to do well not just in her class but in all classes. In the presence of student sorting by teacher quality, these spillover effects will exacerbate the bias in the math teacher quality estimates. Although we cannot separately identify classroom sorting from teacher spillovers,

[29] These rankings have the advantage of potentially greater consistency across tests so long as the reference population of test takers is constant. The publisher of the tests, Riverside Publishing, advertises the TAP as being "fully articulated" with the ITBS and useful for tracking student progress. Less than 2% of the sample is censored, of which over 98% are at the lowest possible percentile score of 1. Estimates using a Tobit to account for this censoring problem result in virtually identical coefficient estimates and estimates of the standard deviation of the $\hat{\tau}_j$.

[30] We have also tried using the robust estimator developed by Huber to account for outliers. The technique weights observations based on an initial regression and is useful for its high degree of efficiency in the face of heavy-tailed data. These results generate an even wider distribution of estimated teacher quality.

Table 10
The Distribution of the Estimated Math Teacher Effects When English Teachers Are Included

	Teacher Quality Estimates		
	Math Only (1)	Math and English (2)	English Only (3)
Math teachers:			
Standard deviation	.273	.278	
Adjusted standard deviation	.193	.170	
Weighted standard deviation	.213	.208	
Number of math teachers	783	783	
English teachers:			
Standard deviation		.257	.254
Adjusted standard deviation		.075	.113
Weighted standard deviation		.208	.209
Number of English teachers		1,049	1,049
p-value, *F*-statistic for math teacher effects	.000	.000	
p-value, *F*-statistic for English teacher effects		.000	.000

NOTE.—See notes to table 6. There are 52,957 students in each specification. Column 1 is the same as col. 5 of table 6. Column 2 additionally includes controls for the English teachers, while col. 3 only controls for English teachers.

we are primarily interested in testing the robustness of our math teacher effects to such controls.

We report estimates that condition on English teachers in table 10. For additional comparison, we also report standard deviations of the English teacher effect estimates both with and without controls for the math teachers. Controlling for English teachers, the math teacher adjusted standard deviation is roughly 0.02 grade equivalents smaller and less precisely estimated. Yet 88% of the math teacher impact remains. However, the size of the English teacher effect is noteworthy on its own. While it is less than half the size (0.075 vs. 0.170) of the dispersion in math teacher quality, it appears to be educationally important. Analogously, when we redo the analysis on reading scores (not reported), the adjusted standard deviation for English teachers is again only slightly smaller, 0.17 versus 0.15 grade equivalents, when we control for other (in this case, math) teachers. Furthermore, the size of the adjusted standard deviation for math teachers is quite notable, roughly 0.12 grade equivalents. Arguably, reading tests are less directly tied to an individual subject. Nevertheless, these results suggest two general conclusions. First, our quality measures, both for math and English teachers, are generally robust to controls for additional teachers. But, second, future research could explore why there are such large achievement effects estimated for teachers whom one would not expect to be the main contributors to a subject area's learning. Can

this be explained by sorting or does a teacher's influence reach beyond his or her own classroom?[31]

G. Heterogeneity by Ability Level, Race, and Sex

Table 11 explores the importance of teachers for different student groups. In columns 1–3, we look at teacher dispersion for students of different "ability." We stratify the sample into ability groups based on the eighth-grade math test score and reestimate the teacher effects within ability group. Low-ability students are defined as those in the bottom one-third of the Chicago public school eighth-grade score distribution, at or below 7.5 grade equivalents. Low-ability students have a mean test score gain of 0.54 grade equivalents. High-ability students are in the top one-third of the eighth-grade test score distribution, with scores above 8.7 (i.e., performing at or above national norms). These students have mean test score gains of 2.2 grade equivalents. All other students are classified as "middle" ability. The middle group has an average gain of 0.67 grade equivalents. Looking at subgroups of students with more similar initial test scores should help reduce the possibility that teacher effect estimates are simply measuring test score growth related to test format and normalization issues. As such, it can be considered another test of the robustness of the results. Moreover, it is of independent interest to document the effect of teachers on different student populations, particularly those achieving at the lowest and highest levels. The major drawback, of course, is that by limiting the sample to a particular subgroup we exacerbate the small sample size problem in estimating teacher quality.

Among all ability groups, we attribute one-third to one-half of the standard deviation in estimated teacher effects to sampling variability. That said, a one standard deviation improvement in teacher quality is still worth a sizable gain in average test score growth: 0.13, 0.20, and 0.13 grade equivalents for low-, middle-, and high-ability students. These outcomes are 24%, 29%, and 6% of average test score gains between eighth and ninth grade for each group, respectively.[32] In relative terms, the largest impact of teachers is felt at the lower end of the initial ability distribution. These results are not sensitive to refinements in the way previous test

[31] As one informal test, we controlled for own student absences to assess whether the mechanism by which English teachers might influence math test scores is to encourage (or discourage) students from attending school. However, we found that own absences have no impact on the dispersion of the English teacher fixed effects.

[32] Although not related directly to the teacher effects, the dynamics of the test scores differ across groups as well. The autoregressive component of math scores is substantially lower for the lowest-achieving students (around 0.47) relative to middle- and high-ability students (1.3 and 1.4).

Table 11
Distribution of the Estimated Teacher Effects for Selected Student Subgroups

	Ability Level			Race/Ethnicity			Sex	
	Low (1)	Middle (2)	High (3)	Non–African American, Non-Hispanic (4)	African American (5)	Hispanic (6)	Male (7)	Female (8)
Mean gain	.54	.67	2.22	2.19	.86	1.19	1.22	1.06
Standard deviation	.236	.304	.274	.259	.293	.248	.303	.264
Adjusted standard deviation	.129	.196	.132	.105	.201	.132	.201	.160
p-value, *F*-statistic for teacher effects	.000	.000	.000	.003	.000	.000	.000	.000
Number of teachers	518	478	390	204	579	353	627	620
Number of students	16,880	18,616	17,461	6,940	29,750	16,271	25,299	27,658

NOTE.—See notes to table 6. Ability level is assigned in thirds based on the eighth-grade test score distribution. High-ability students have scores above 8.7, middle-ability students have scores between 7.5 and 8.7, and low-ability students have scores of less than 7.5. All regressions include school fixed effects and the student, peer, and neighborhood covariates included in the table 6, cols. 3 and 5, specifications.

score results are controlled, including allowing for nonlinearities in the eighth-grade score or controlling for sixth- and seventh-grade scores.

By race, teachers are relatively more important for African American and, to a lesser extent, Hispanic students. A one standard deviation, one semester increase in teacher quality raises ninth-grade test score performance by 0.20 grade equivalents (23% of the average annual gain) for African American students and 0.13 grade equivalents (11% of the average annual gain) for Hispanic students. The difference is less important for non–African American, non-Hispanic students both because their mean test score gain is higher and because the estimated variance in teacher effects is somewhat smaller.

There is very little difference in the estimated importance of teachers when we look at boys and girls separately. The adjusted standard deviation of teacher effects equals 0.20 for boys and 0.16 for girls. For both girls and boys, a one standard deviation improvement in teacher quality translates into a test score gain equal to 15%–16% of their respective average annual gains.

Finally, we examined whether quality varies within teacher depending on the initial ability of the student. That is, are teachers that are most successful with low-ability students also more successful with their high-ability peers? To examine this issue, we use the 382 math teachers in our sample that have at least 15 students in both the top half and bottom half of the eighth-grade math test score distribution. We then explored whether teachers ranked in the bottom (or top) half of the quality rankings when using low-ability students are also ranked in the bottom (or top) half of the ability distribution when using high-ability students. We find that 67% of low-ranking teachers for low-ability students are low-ranking teachers for high-ability students. Sixty-one percent of those teachers ranked in the top half using low-ability students are ranked similarly for high-ability students. The correlation between the teacher quality estimates derived from low- and high-ability teachers is a highly statistically significant 0.39, despite small sample sizes that accentuate sampling error. Therefore, there is some evidence that teacher value added is not specific to certain student types; a good teacher performs well, for example, among both low- and high-ability students.

V. Predicting Teacher Quality Based on Resume Characteristics

This final section relates our estimates of τ_j to measurable characteristics of the instructors available in the CPS administrative records. Observable teacher characteristics include demographic and human capital measures such as sex, race, potential experience, tenure at the CPS, advanced degrees (master's or PhD), undergraduate major, undergraduate college attended,

and teaching certifications.[33] We report select results in table 12. All are based on the full control specification reported in column 5 of table 6. We discuss common themes below.

First and foremost, the vast majority of the total variation in teacher quality is unexplained by observable teacher characteristics. For example, a polynomial in tenure and indicators for advanced degrees and teaching certifications explain at most 1% of the total variation, adjusting for the share of total variation due to sampling error.[34] That is, the characteristics on which compensation is based have extremely little power in explaining teacher quality dispersion. Including other teacher characteristics, changing the specifications for computing the teacher effects, and altering the minimum student-semester threshold have little impact on this result. In all cases, the R^2 never exceeds 0.08.

Given a lack of compelling explanatory power, it is of little surprise that few human capital regressors are associated with teacher quality.[35] Standard education background characteristics, including certification, advanced degrees, quality of college attended, and undergraduate major, are loosely, if at all, related to estimated teacher quality. Experience and tenure

[33] Potential experience is defined as age − education − 6 and is averaged over the 3 years of the sample.

[34] The R^2 is an understatement of the explanatory power since a significant fraction, perhaps up to a third, of the variation in $\hat{\tau}_j$ is due to sampling error. If we simply multiply the total sum of squares by a rather conservative 50% to account for sampling variation, the R^2 will double. However, in all cases it is never higher than about 15%. By comparison, the R^2 from a wage regression with education, experience, gender, and race using the 1996–99 Current Population Survey is about 0.2, without any corrections for sampling variation. Furthermore, firm-specific data or modeling unobserved person heterogeneity causes the R^2 on productivity and wage regressions to be quite a bit higher (e.g., Abowd, Kramarz, and Margolis 1999; Lazear 1999).

[35] Other studies that correlate specific human capital measures to teacher quality are mixed. Hanushek (1971) finds no relationship between teacher quality and experience or master's degree attainment. Rivkin et al. (2005) also find no link between education level and teacher quality, although they find a small positive relationship between the first 2 years of teacher experience and teacher quality. Kane et al. (2006) find a positive experience effect in the first few years as well. Summers and Wolfe (1977) find that student achievement is positively related to the teacher's undergraduate college while student achievement is negatively related to the teacher's test score on the National Teacher Examination test. In contrast, Hanushek (1971) finds that teacher verbal ability is positively related to student achievement for students from "blue-collar" families. Ferguson (1998) argues that teacher test score performance is the most important predictor of a teacher's ability to raise student achievement. Goldhaber and Brewer (1997) find some evidence that teacher certification in mathematics or majoring in mathematics is positively related to teacher quality, but Kane et al.'s (2006) results suggest otherwise. Other work on teacher training programs is likewise mixed (e.g., Angrist and Lavy 2001; Jacob and Lefgren 2004).

Table 12
Impact of Observable Characteristics on Teacher Fixed Effects

	(1)	(2)	(3)
Female		.073*	.069*
		(.020)	(.020)
Asian		.007	.008
		(.041)	(.041)
Black		.050*	.048*
		(.023)	(.023)
Hispanic		−.057	−.056
		(.039)	(.039)
Potential experience		.004	
		(.008)	
Squared		.000	
		(.000)	
Cubed (divided by 1,000)		.004	
		(.007)	
Potential experience<= 1:			.021
			(.042)
Master's	.002	.004	.007
	(.020)	(.020)	(.020)
PhD	−.103	−.077	−.068
	(.077)	(.076)	(.076)
BA major: education	.003	−.012	−.016
	(.030)	(.034)	(.033)
BA major: math	.003	.022	.021
	(.024)	(.025)	(.025)
BA major: science	.001	.029	.035
	(.040)	(.040)	(.040)
Certificate, bilingual education		−.067*	−.069*
		(.037)	(.037)
Certificate, child		.121	.120
		(.082)	(.082)
Certificate, elementary		.004	.006
		(.038)	(.038)
Certificate, high school		−.033	−.033
		(.033)	(.032)
Certificate, special education		.007	.008
		(.037)	(.036)
Certificate, substitute		−.004	−.005
		(.026)	(.026)
Tenure at CPS	−.001	−.001	.003
	(.008)	(.010)	(.009)
Squared	.000	.000	.000
	(.001)	(.001)	(.001)
Cubed (divided by 1,000)	.004	.005	.009
	(.011)	(.012)	(.011)
BA university, *US News* 1		−.010	−.014
		(.037)	(.037)
BA university, *US News* 2		.013	.012
		(.037)	(.037)
BA university, *US News* 3		.004	.002
		(.029)	(.029)
BA university, *US News* 4		.003	.003
		(.038)	(.038)
BA university, *US News* 5		−.003	.002
		(.072)	(.072)
BA university, local		.008	.005
		(.023)	(.022)
Adjusted R^2	.005	.077	.074
Number of teachers with observables	589	589	589

NOTE.—The dependent variable is teacher quality estimated using the table 6, col. 5, specification. Each specification also includes a constant. Potential experience is calculated as age − education − 6 and is the teacher's average over the 3 years.

* Significant at 10% level.

have little relation to τ_j when introduced in levels (unreported), higher order polynomials (col. 2), or as a discontinuous effect of rookie teachers (col. 3). We have also tried identifying experience and/or tenure effects from a specification that includes teacher-year fixed effects (rather than just teacher fixed effects) which allows us to use variation within teacher over time, using various combinations of intervals for experience and tenure (e.g., 0–3, 3–7, 7–10, 10 plus), and capping experience at 10 years. None of these adjustments show a large or statistically important effect for either tenure or experience. Rather, at best, it appears that there is a 0.02 grade-equivalent increase in quality over the first few years of experience that flattens and eventually recedes. Given our sample sizes, such an effect is impossible to precisely estimate.

Female and African American teachers are associated with test scores roughly 0.07 and 0.05 grade equivalents higher than male and white teachers. Some of this influence derives from students with similar demographics.[36] In particular, African American boys and girls increase math test scores by 0.067 (standard error of 0.037) and 0.042 (standard error of 0.034) grade equivalents in classrooms with an African American teacher rather than a white teacher. However, we do not find an analogous result for Hispanic student-teacher relationships. Across all student race groups, including Hispanics, math test scores are 0.05–0.10 grade equivalents lower in classrooms with Hispanic teachers.

Likewise, female teachers have a larger impact on female students, especially African Americans. African American girls increase math test scores by 0.066 (standard error of 0.032) grade equivalents when in a classroom with a female teacher. This compares to a 0.032 (standard error of 0.033) grade equivalent boost for boys. Because of small sample sizes, we cannot distinguish Hispanic boys from Hispanic girls, but among all Hispanic students, female teachers boost math test scores by 0.060 (standard error of 0.024) grade equivalents. All of these results are similar under simpler specifications that include only the race and/or gender of the teacher.

VI. Conclusion

The primary implication of our results is that teachers matter. While this has been obvious to those working in the school systems, it is only in the last decade that social scientists have had access to data necessary to verify and estimate the magnitude of these effects. In spite of the improved data, the literature remains somewhat in the dark about what

[36] Goldhaber and Brewer (1997) find teacher quality higher among female and lower among African American instructors. Ehrenberg, Goldhaber, and Brewer (1995) and Dee (2004) also look at teacher race and/or sex but instead focus on whether students perform better with teachers of their own race and/or sex.

makes a good teacher. Our results are consistent with related studies like Hanushek (1992) and Rivkin et al. (2005), who argue that characteristics that are not easily observable in administrative data are driving much of the dispersion in teacher quality. Traditional human capital measures have few robust associations with teacher quality and explain a very small fraction of its wide dispersion. That our teacher quality measure persists over time implies that principals may eventually be able to identify quality; however, they are unlikely to have information on teacher quality when recruiting or for recent hires for whom little or no information is available on the teacher's effect on students' test score achievement. More generally, teacher quality rankings can be quite sensitive in a value-added framework when across-school differences are ignored. Without such controls, naive application of value added may undermine teacher performance incentives. One common proposal is to tie teacher pay more directly to performance, rather than the current system, which is based on measures that are unrelated to student achievement, namely, teacher education and tenure. That said, such a compensation scheme would require serious attention to implementation problems (Murnane et al. 1991), including, but far from limited to, important measurement issues associated with identifying quality.

Data Appendix

The student administrative records assign an eight-character identification code to teachers. The first three characters are derived from the teacher's name (often the first three characters of the last name) and the latter five reflect the teacher's "position number," which is not necessarily unique. In the administrative student data, several teacher codes arise implausibly few times. When we can reasonably determine that the teacher code contains simple typographical errors, we recode it in the student data. Typically, we will observe identical teacher codes for all but a few students in the same classroom, during the same period, in the same semester, taking the same subject, and a course level other than special education. These cases we assume are typographical errors. Indeed, often the errors are quite obvious, as in the reversal of two numbers in the position code.

A second problem we face in the teacher data occurs because a teacher's position and school number may change over time. We assume that administrative teacher records with the same first and last name and birth date are the same teacher and adjust accordingly. Additionally, for position numbers that appear to change over time in the student data, we made assumptions about whether it was likely to be the same teacher based on the presence of the teacher in that school in a particular year in the teacher administrative data.

Finally, we match students to teachers using a three-letter name code and the position number for the combinations that are unique in the teacher data.[37]

References

Abowd, John M., Francis Kramarz, and David Margolis. 1999. High wage workers and high wage firms. *Econometrica* 67, no. 2:251–333.

Angrist, Joshua D., and Victor Lavy. 2001. Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics* 19, no. 2:343–69.

Borjas, George J. 1987. Self-selection and the earnings of immigrants. *American Economic Review* 77, no. 4:531–53.

Borjas, George J., and Glenn T. Sueyoshi. 1994. A two-stage estimator for probit models with structural group effects. *Journal of Econometrics* 64, no. 1–2:165–82.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2004. Teacher sorting, teacher shopping, and the assessment of teacher effectiveness. Unpublished manuscript, Public Policy Studies, Duke University.

Coleman, James S., et al. 1966. *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Dee, Thomas S. 2004. Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics* 86, no. 1: 195–210.

Ehrenberg, Ronald G., Daniel D. Goldhaber, and Dominic J. Brewer. 1995. Do teachers' race, gender, and ethnicity matter? *Industrial and Labor Relations Review* 48, no. 3:547–61.

Ferguson, Ronald. 1998. Paying for public education. *Harvard Journal of Legislation* 28:465–98.

Goldhaber, Dan D., and Dominic J. Brewer. 1997. Why don't school and teachers seem to matter? *Journal of Human Resources* 32, no. 3:505–23.

Greenwald, Rob, Larry Hedges, and Richard Laine. 1996. The effect of school resources on student achievement. *Review of Educational Research* 66:361–96.

Grogger, Jeff, and Eric Eide. 1995. Changes in college skills and the rise in the college wage premium. *Journal of Human Resources* 30, no. 2: 280–310.

Hanushek, Eric A. 1971. Teacher characteristics and gains in student achievement. *American Economic Review* 61, no. 2:280–88.

———. 1992. The trade-off between child quantity and quality. *Journal of Political Economy* 100, no. 1:84–117.

[37] Note that we assigned some three-letter teacher codes for cases in which the teacher code did not correspond to the first three letters of the teacher's last name.

———. 1996. Measuring investment in education. *Journal of Economic Perspectives* 10, no. 4:9–30.

———. 1997. Assessing the effects of school resources on student performance: An update. *Education Evaluation and Policy Analysis* 19: 141–64.

———. 2002. Publicly provided education. In *Handbook of public finance*, vol. 4, ed. Alan Auerbach and Martin Feldstein. Amsterdam: North-Holland Press.

Hanushek, Eric A., and Dennis D. Kimko. 2000. Schooling, labor-force quality, and the growth of nations. *American Economic Review* 90, no. 5:1184–1208.

Hoxby, Caroline. 2000. Peer effects in the classroom: Learning from gender and race variation. Working paper no. 7867, National Bureau of Economic Research, Cambridge, MA.

Jacob, Brian A., and Lars Lefgren. 2004. The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources* 39, no. 1: 50–79.

Jacob, Brian A., and Steven D. Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118, no. 3:843–77.

Jepsen, Christopher, and Steven Rivkin. 2002. What is the tradeoff between smaller classes and teacher quality? Working paper no. 9205, National Bureau of Economic Research, Cambridge, MA.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2006. What does certification tell us about teacher effectiveness? Evidence from New York City. Working paper no. 12155, National Bureau of Economic Research, Cambridge, MA.

Kane, Thomas J., and Douglas O. Staiger. 2002. The promises and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16, no. 4:91–114.

———. 2005. Using imperfect information to identify effective teachers. Working paper, Department of Economics, Dartmouth University.

Lazear, Edward. 1999. Personnel economics: Past lessons and future directions: Presidential address to the Society of Labor Economists, San Francisco, May 1, 1998. *Journal of Labor Economics* 17, no. 2:199–236.

Manski, Charles F. 1993. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* 60, no. 3:531–42.

Moulton, Brent. 1986. Random group effects and the precision of regression estimates. *Journal of Econometrics* 32:385–97.

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

Murnane, Richard, Judith Singer, John Willett, James Kemple, and Randall Olsen. 1991. *Who will teach? Policies that matter*. Cambridge, MA: Harvard University Press.

Rivers, June, and William Sanders. 2002. Teacher quality and equity in educational opportunity: Findings and policy implications. In *Teacher quality*, ed. Lance T. Izumi and Williamson M. Evers. Stanford, CA: Hoover Institution Press.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73, no. 2:417–58.

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94, no. 2:247–52.

Sacerdote, Bruce. 2001. Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics* 116, no. 2: 681–704.

Summers, Anita A., and Barbara L. Wolfe. 1977. Do schools make a difference? *American Economic Review* 67, no. 4:639–52.

U.S. Census Bureau. Census 2000, summary file 1. Generated by authors using American FactFinder, http://factfinder.census.gov, accessed October 23, 2006.

U.S. Department of Education. National Center for Education Statistics. 2000. *The condition of education 2000*. NCES publication no. 2000–062. Washington, DC: U.S. Government Printing Office.

U.S. Department of Education. National Center for Education Statistics. 2003. *Characteristics of the 100 largest public elementary and secondary school districts in the United States: 2001–02*. NCES 2003-353, Jennifer Sable and Beth Aronstamm Young. Washington, DC: U.S. Government Printing Office.

U.S. News & World Report. 1995. *America's best colleges*. Washington, DC: U.S. News & World Report.

Copyright of Journal of Labor Economics is the property of University of Chicago Press and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

EXHIBIT 6
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

EDUCATION**SECTOR** REPORTS

May 2008

WAITING TO BE WON OVER:

Teachers Speak on the Profession, Unions, and Reform

By Ann Duffett, Steve Farkas, Andrew J. Rotherham, and Elena Silva

TABLE OF CONTENTS

Acknowledgements v

Introduction 1

Systemic Problems 2

Considering Change 4

Unions as Protectors and Reformers 8

Newcomers and Veterans 12

Conclusion 16

Appendix A: Methodology 17

Appendix B: Survey 18

© Copyright 2008 Education Sector.

Education Sector encourages the free use, reproduction, and distribution of our ideas, perspectives, and analysis. Our Creative Commons licensing allows for the noncommercial use of all Education Sector authored or commissioned materials. We require attribution for all use. For more information and instructions on the commercial use of our materials, please visit our Web site, www.educationsector.org.

Cover photo courtesy of Liliboas/iStockphoto.

1201 Connecticut Ave., N.W., Suite 850, Washington, D.C. 20036
202.552.2840 • www.educationsector.org

ACKNOWLEDGEMENTS

The Joyce Foundation provided funding for this project. The findings and conclusions are those of the authors alone and do not necessarily represent the opinions of the foundation.

The authors would like to thank all of the teachers who participated in the focus groups and completed the survey.

ABOUT THE AUTHORS

ANN DUFFETT and **STEVE FARKAS** are the two principals of FDR Group, a nonpartisan public opinion research firm (www.thefdrgroup.com). They have conducted dozens of surveys on education, including *Stand By Me: What Teachers Really Think about Unions, Merit Pay and Other Professional Matters* (Public Agenda, 2003).

ANDREW J. ROTHERHAM is co-founder and co-director of Education Sector and a member of the Virginia Board of Education. He is also on the board of directors of the National Council on Teacher Quality. He can be reached at arotherham@educationsector.org.

ELENA SILVA is a senior policy analyst at Education Sector, where she oversees the organization's teacher quality work. She can be reached at esilva@educationsector.org.

ABOUT EDUCATION SECTOR

Education Sector is an independent think tank that challenges conventional thinking in education policy. We are a nonprofit, nonpartisan organization committed to achieving real, measurable impact in education, both by improving existing reform initiatives and by developing new, innovative solutions to our nation's most pressing education problems.

American public education is in the midst of intense change, and teachers, in particular, are facing pressure to produce better outcomes for students. As policymakers, teachers unions, and other stakeholders react to changing demands on the nation's public education system, there remains considerable debate about what teachers think and what they want. Too often assumptions define the conversation rather than actual evidence of teachers' views. Teachers unions and associations often claim to represent the voice of all teachers. But, while these groups serve an important role, they cannot possibly be expected to represent the diverse viewpoints of a profession with 3.2 million practitioners.[1] As such, independent public opinion research that investigates what teachers think about various issues is a necessary contribution to the national conversation on education policy and reform.

In an effort to facilitate and inform this conversation, Education Sector and the FDR Group surveyed 1,010 K–12 public school teachers about their views on the teaching profession, teachers unions, and a host of reforms aimed at improving teacher quality.[2] The survey asks specific questions about the work teachers do and about reform proposals that are currently being debated. It also examines the views of new teachers and veterans. And, when possible, the survey discerns trends by asking some identical questions from a 2003 national survey of K–12 public school teachers and comparing the responses.[3] In order to probe themes and develop the survey instrument, Education Sector and the FDR Group conducted six focus groups in five cities with approximately 60 current public school teachers. (See appendices for a full discussion of methods and the questionnaire.)

The survey revealed that it is hard to place teachers definitively in any one camp even though advocates on all sides of various issues do just that. As a whole, teachers today are what political analysts might describe as "in play" and waiting to be won over by one side or another. Despite frustrations with schools, school districts, their unions, and a number of aspects of the job in general, teachers are not sold on any one reform agenda. They want change but are a skeptical audience. For instance, nearly half of teachers surveyed say that they personally know a teacher who is ineffective and should not be in the classroom. But, although teachers want something done about low-performing colleagues, they are leery of proposals to substantially change how teachers can be dismissed.

Today's teachers have different expectations than teachers in the past, and they expect different things from their professional lives. Yet, they recognize the problems that undermine their profession, including job lock, weak evaluation and reward structures, and too much bureaucracy. With reformers pushing hard for change and teachers unions holding tight to tradition, teachers are caught in the middle, unsure of how their profession should change but very aware that it needs to.

Teachers see problems with their unions as well. For example, many say that the union sometimes fights to protect teachers who really should be out of the

---

[1] Digest of Educational Statistics, 2007.

[2] The term "union" is used throughout this report to refer to unions and associations. The survey referenced both.

[3] Steve Farkas, Jean Johnson, and Ann Duffett, *Stand By Me: What Teachers Really Think About Unions, Merit Pay and Other Professional Matters*, (New York: Public Agenda, 2003).

classroom. But teachers still see the union as essential, and they value the union's traditional role in safeguarding their jobs. New teachers are more likely today than they were in 2003 to call unions "absolutely essential." And many teachers would like to see their unions explore some new activities, especially some of the ideas associated with the "new unionism" agenda, and take the greater role in reform, but not if that comes at the expense of the union's core mission.

The fluid environment presents both challenges and opportunities for education leaders and policymakers. Teachers unions may claim a deep loyalty from their members but the relationship seems to be based mostly on the practical benefits that the union provides. Likewise, school districts face high hurdles to convince teachers that they have their best interests in mind and deserve their trust. And in an environment of distrust, reformers face real challenges to earn the support of teachers and turn today's most popular reform ideas aimed at improving teaching and learning into public policy.

This report is organized into four sections. The first highlights key findings about the challenges that teachers see in their profession, including weak evaluation processes and a rigid tenure and pay system. The second section describes how teachers feel about a range of reforms aimed at improving their profession, from new evaluation approaches to differential pay proposals. The third section focuses on teachers' opinions about their union and what they feel the union role should be in improving teacher quality. The final section examines some key points of comparison between new teachers, who have been on the job fewer than five years, and veteran teachers, who have been teaching for more than 20 years.

SYSTEMIC PROBLEMS

In order for teacher quality to improve, there are some systemic problems in the profession that must be changed. Teachers, for instance, say the benefits structure works against teacher quality by locking in people who would rather move on or retire, and laws and contractual rules hinder quality by making it difficult to remove persistently ineffective teachers. Teachers also point to weak evaluation procedures and bureaucracy as serious problems that hold back the profession.

Locked In

Three in four public school teachers (76 percent) agree that, "Too many veteran teachers who are burned out stay because they do not want to walk away from the benefits and service time they have accrued." And this view resonates with majorities of teachers whether they are newcomers to the profession (80 percent) or veterans (68 percent). [Fig. 1–1]

Figure 1–1.* Trapped by Benefits

Too many veteran teachers who are burned out stay because they do not want to walk away from the benefits and service time they have accrued. [Question 8]

*Percentages in figures may not equal 100 percent due to rounding or omission of answer categories. Question wording may be edited for space. Full question wording is available in Appendix B. Small discrepancies between percentages in the text and figures are due to rounding.

According to teachers, not only do the system's incentives lock in teachers who'd rather leave; its rules make it hard to push colleagues out when they really should not be teaching. Well over half of the teachers surveyed (55 percent) say that in their district it is very difficult and time-consuming to remove clearly ineffective teachers who shouldn't be in the classroom but who are past their probationary period. Only 13 percent say this is not the case. And almost half of teachers (46 percent) say they know a teacher in their own building who is past the probationary period but who is clearly ineffective and shouldn't be in the classroom (42 percent say they do not know such a teacher). [Fig. 1–2]

Figure 1–2. Breaking Up Is Hard to Do

In some districts, the process for removing teachers who are clearly ineffective and shouldn't be in the classroom—but who are past the probationary period—is very difficult and time-consuming. Is this the case in your district, or not? [Question 67]

At the same time, teachers can't point to a single, preferred strategy that principals use to deal with teachers who clearly should not be in the classroom. One in four teachers (26 percent) says a principal in their district, if faced with a persistently ineffective teacher, would "make a serious effort to retrain the teacher." And 18 percent say a principal in their district would most likely "initiate formal proceedings to remove the teacher." But some think that their district's principals would be most likely to do nothing (14 percent); or that they would "transfer the teacher to another school" (14 percent). Another 13 percent say the principals would "quietly encourage the teacher to leave." [Fig. 1–3]

Figure 1–3. No Clear Solution to Ineffectual Teachers

Which is the most likely course of action a principal in your district would take if faced with a persistently ineffective teacher who was already past the probationary period? [Question 70]

Weak Evaluations

Teachers indicate that the most obvious technique used to assess teacher quality—the formal observation and evaluation—is not doing the job. In fact, only 26 percent of teachers report that their own most recent formal evaluation was "useful and effective." The plurality—41 percent—say it was "just a formality," while another 32 percent say at best it was "well-intentioned but not particularly helpful" to their teaching practice. Almost seven in 10 teachers (69 percent) say that when they hear a teacher at their school has been awarded tenure, they think that it's "just a formality—it has very little to do with whether a teacher is good or not." [Fig. 1–4]

Figure 1–4. Evaluations: Just a Formality

Which statement describes your most recent formal evaluation? [Question 28]

Voices From the Field …

On Benefits:

An experienced teacher in Chicago described the calculations going through her mind as she approaches retirement: *"They will take 5 percent of your pension away from you for every year that you quit before the age of 60. If you're at 30 years, and you're burned out, you better go the 34, or they're going to take 20 percent of your pension from you. That is the really bad thing about it because it almost makes these teachers be there whether they want to be or not."*

Newcomers talked about being forewarned about the pitfalls of the benefit structure: *"I've been around four years, and I've heard people say, 'If you want to get out of the system, get out of it now before you're locked in,'"* explained a relatively novice Milwaukee teacher.

On Ineffective Peers:

Teachers acknowledged the existence of ineffective peers and how hard it is to remove them from the classroom. *"I have children in school right now, and there certainly are some teachers that I will not let my children go into their classrooms,"* said a teacher working in the suburbs of Milwaukee.

A teacher from the Milwaukee public schools described her school's effort to remove a problematic colleague: *"They have to go through a lot of hoops. … We had blatant documentation, parent complaints, calls to a school board, all sorts of things, but the principal's hands were tied on every single situation."*

And a teacher from a Milwaukee suburb said: *"In our district there's a male teacher. … He is struggling very, very much and is still probationary, but they renewed his contract. Our entire department is shocked."*

Few Rewards

Outstanding teachers are unlikely to be recognized in any formal way, if at all. Half of teachers (49 percent) say school and district officials "do not reward outstanding teachers; the reward is solely intrinsic." Twenty-nine percent say outstanding teachers receive "official recognition (for example, formal commendation or note to file)," 16 percent say they receive some form of "informal recognition (for example, better treatment or perks)," and 10 percent say they get a "token gift."

Only 5 percent say that teachers receive a "financial bonus" for outstanding work. Ninety-seven percent of teachers say salary increases in their district are determined "according to a strictly defined schedule mostly driven by their years of service and the credits they attain."

A Lot of Rules

Teachers describe schools that are tied up by bureaucracy, governed by a clutter of rules, legal stipulations, and contractual obligations that force good teaching and learning to take a back seat. The vast majority of teachers (86 percent) agree that, "Teachers are required to do too much paperwork and documentation about what goes on in their classrooms."

They acknowledge that principals also work under difficult conditions, strapped for time and bogged down by the restrictions of a heavily bureaucratic system. Almost six in 10 (59 percent) agree that, "All the paperwork and legal and contractual restrictions make it difficult for principals to get things done"; only 28 percent disagree. In the end, when things go wrong, teachers are somewhat more likely to blame the system than to blame principals. More than half (55 percent) *reject* the view that, "When individual schools fail, it's usually because they have ineffective principals at the helm," although 40 percent agree with that statement.

Most teachers, by a 68 percent to 21 percent margin, say schools would be better for students if principals and teachers had more control and flexibility about work rules and school duties. [Fig. 1–5]

Figure 1–5. More Flexible Schools

On the whole, which type of school would be better for students? One where:
[Question 66]

Summary and Analysis

Public school teachers in this national survey depict a system that seems to be stuck when it comes to fine-tuning its workforce and making the most of its professionals' talents. Teachers who would rather move on are often trapped by benefits, and teachers who *should* move on are often unduly protected. "We do have people in the profession that get there and are entrenched, burned out. I remember there was one teacher who wanted to laminate her lesson plans. There are those people, and that's a negative stigma against our profession," said a Phoenix teacher.

Yet, even when teachers are identified as not being effective, the system does not make it easy to get rid of them, primarily because the most common technique used to assess teacher quality and award tenure—the formal observation and evaluation—is not doing the job. When evaluations are a mere formality, as many teachers say they are, not only do teachers lose out on the chance to improve their craft, but ineffective teachers slip through and gain tenure.

According to these findings, teachers see themselves, and the principals who lead them, as overly constrained by work rules that define what they can do, how they should do it, and when it can be done. As a result, they feel treated as less than professional. "It's demoralizing," said one New York City teacher about having to punch a timecard each day. "I have a master's degree plus 30 credits, but I have a timecard with my name on it. … It's ridiculous."

The findings suggest support among teachers for a system that has more flexible work rules, more trust in teachers' judgment and professionalism, and where decisions about teacher quality are not dependent on rigid rules, weak evaluations, and faulty tenure systems.

CONSIDERING CHANGE

In order to change and eliminate systemic problems, school systems will need the support of teachers. Many public school teachers are open to some new ways of evaluating, rewarding, and paying teachers as well as ideas for attracting and retaining high-quality teachers. But some proposals do not gain wide support. For instance, teachers are resistant to using student test scores as a way of measuring teacher effectiveness in the classroom, and most oppose the idea of offering higher starting salaries in exchange for smaller pensions.

Stronger Evaluations

Concerned that the current evaluation process is weak and often no more than a formality, teachers express a willingness to reform the tenure system. Almost eight in 10 teachers (79 percent) support strengthening the formal evaluation of *probationary* teachers so that they will get tenure only after they've proven to be very good at what they do. Tenured teachers are more likely to support this proposal than their non-tenured colleagues (83 percent vs.

66 percent). What's more, most teachers think that even *tenured* teachers should be formally evaluated on a regular basis. Evaluations should occur each year according to 31 percent of tenured and 36 percent of non-tenured teachers, and at least every two years according to 26 percent of tenured and 30 percent of non-tenured teachers. [Fig. 2–1]

Figure 2–1. Regular Evaluations for Tenured Teachers

*Statistically significant difference.

Putting Pay on the Table

Teachers show support for some pay proposals, especially those that reward demanding assignments or additional work.

Eighty percent of public school teachers favor giving financial incentives to "teachers who work in tough neighborhoods with low-performing schools," an increase of 10 percentage points from the 70 percent of teachers who favored an identical proposal in 2003. A large majority of teachers (64 percent) also favor giving financial incentives to "teachers who have pursued and achieved accreditation from the National Board for Professional Teaching Standards," an increase of 7 percentage points since the question was asked in 2003.

More than half of teachers (53 percent) favor giving financial incentives to "teachers who specialize in hard-to-fill subjects such as science or mathematics," an increase of 11 percentage points from 2003. And a solid majority (58 percent) favors giving financial incentives to "teachers who consistently receive outstanding evaluations by their principals." [Fig. 2–2]

Even a proposal to trade tenure or job protection for higher pay garners some support—although it is hardly overwhelming. One in four tenured teachers (25 percent) would trade their tenure for a pay increase of $5,000 per year, while an additional 29 percent would consider the trade if the pay increase was "a lot higher." About three in 10 (29 percent) reject the idea outright. [Fig. 2-3]

Figure 2–2. What Merits More Pay?

*Statistically significant difference.

Figure 2–3. Swapping Tenure

*Statistically significant difference.

Still Uneasy About Test Scores

Teachers are resistant to using test scores as a measurement of their performance and pay. As shown in Figure 2-2, one in three teachers (34 percent) favors giving financial incentives to teachers "whose kids routinely score higher than similar students on standardized tests." Most teachers today (64 percent) oppose the idea, up 8 percentage points from the 56 percent who opposed it in 2003.

Almost half of the public school teachers surveyed (49 percent) say it's an excellent (15 percent) or good (34 percent) idea to measure teacher effectiveness based on student growth, or "to assess students' skills and knowledge when they first come to a teacher and to measure them again when students leave." But almost half (48 percent) say it is a poor or fair idea. Similarly, the percent of public school teachers who favor (44 percent) or oppose (51 percent) financially rewarding teachers whose students make comparatively more academic progress in terms of "improved reading levels, teacher evaluations, and classroom tests" hovers around the halfway mark. [Fig. 2–4]

Figure 2–4. Adding Value

Attracting and Retaining Teachers

Teachers show strong support for recruitment strategies that improve the conditions and flexibility of their work. The majority of teachers (85 percent) agree that it is an excellent or good idea to "give teachers more time during the school day for class preparation and planning" as a way to attract and retain high-quality teachers. Almost eight in 10 (78 percent) say it is either an excellent or good idea to "make it far easier to leave and return to teaching without losing retirement benefits." And seven in 10 (70 percent) think positively of a proposal that would "make it easier to earn and take sabbatical leave for teachers working in really challenging schools."

Support among public school teachers is also apparent for recruiting teachers from other industries and easing state certification requirements. Fifty-four percent say they would be open to "actively recruiting and training second-career candidates from business and other sectors of the economy." High school teachers are more likely than elementary school teachers to support this idea (62 percent vs. 49 percent). However, only 43 percent of teachers think it is an excellent or good idea "to ease state certification requirements while intensifying in-class supervision, observation, and mentoring."

Teachers are less supportive about using financial incentives to attract teachers. Less than half (43 percent) of teachers say it is an excellent or good idea to "pay signing bonuses to attract new talent." Non-union members are more likely than union members to be supportive of this proposal (52 percent vs. 39 percent). And minority teachers (self-identified as African-American, Hispanic, or Asian) are more likely to support this idea than white teachers (63 percent vs. 41 percent).

Only 22 percent of public school teachers think it's an excellent or good idea to offer new teachers "substantially higher starting salaries in exchange for smaller pensions when they retire." A large majority (71 percent) view it as either a poor or fair idea. Large majorities of both union members and non-members view this proposal negatively (72 percent of members and 69 percent of non-members). [Fig. 2–5]

Figure 2–5. Attracting the Best and Brightest

†Figures in parentheses represent totals.

Summary and Analysis

This national survey highlights public school teachers' willingness to explore new ideas for assessing and rewarding performance and for attracting new candidates to the profession. And while teachers overall are far from convinced that differences in teacher impact can be measured fairly or measured at all, there does seem to be room for negotiation around pay initiatives, an insight that is often masked in the highly charged debates about pay for performance proposals. Still, some teachers may be insulted by the very idea that they would respond to financial incentives or that public schools would benefit from the management techniques used in the private sector.

A series of survey questions on differentiated pay approaches demonstrates there are situations where teachers agree that paying colleagues differently is justified. Teachers know that working in tough neighborhoods with low-performing schools is a difficult

assignment and may feel that it's only fair to reward those willing to put forth the extra effort. Teachers also feel positively about paying more money to colleagues who have pursued and achieved National Board certification. Many know that it takes a lot of work to get such accreditation and may feel that this extra effort deserves reward. Most likely, teachers favor bonus pay for tough assignments and National Board certification because they are familiar with these ideas and know colleagues who have benefited from them. Teachers may also be amenable to these ideas because they see them as "objective" and not susceptible to favoritism. There has been growing support among teachers for some of these proposals; three of the five proposals about financial incentives that we tested have gained support in the past four years.

Paying teachers based on student test scores, however, remains controversial. Teachers' suspicion of standardized tests as a fair and objective measure drives some of this resistance. "To reward teachers for great test scores is absurd," commented one teacher. "There is such a range of external issues that work in a classroom; there is no way to accurately assess how great a job [a teacher] is doing based on test scores! And if it is based on test scores, who ultimately decides? How can favoritism, cronyism, and all other matters of human subjectiveness not come into play?"

Teachers also appear to be split over the use of growth or "value-added" measures of teacher performance, which assess teacher effectiveness based on student progress over time. Despite increased attention to these ideas in education policy, teachers are no more likely today to support them than they were in 2003.

Even though teachers show openness to some changes, it's hard to ignore the misgivings they revealed (across different types of questions) about using student achievement to measure, evaluate, and compensate them. And in focus groups, teachers bristled at even the suggestion that they should be solely responsible for a child's academic achievement when so many others—parents, administrators, even the students themselves—are not doing their part. In Milwaukee, a city school teacher told us, "I would love to be rewarded for the merits that I do make, but I would not like to be penalized for things that are out of my control."

Voices From the Field …

On Merit Pay:
A Chicago teacher's experience illustrates the mixed feelings teachers may have on value-added measures, especially when used to determine pay bonuses: *"Prior to this year, I would say that merit-based pay … it's an insult. Like I'm going to work harder? I work as hard as I can, and I'm not going to work harder for more money. That's an insult to me. My school … got a very large federal grant starting next year for this merit-based pay, so the way that it's worked out, I really like it. … It's based on the value-added of what I do. … My kids started at 10. … So if I take that from a 10 to a 25, are they at grade level? No. Did I do a really good job? Did I bring them up significantly? Yes. … We'll see. It's not been implemented yet. In theory, I like it."*

To teachers, one of the critical downsides of differentiating teacher pay—whatever the approach—is that it will breed unhealthy competition and wreak havoc on the collaborative spirit that they see as essential to effective teaching and student learning. One teacher wrote, *"I still feel that teaching is one of the most valuable and fulfilling professions in the world. I am afraid that by tying teacher compensation to effectiveness, there will be less willingness for teachers to work together. Teaching will become a competition for getting the most money."*

A New York teacher said, *"Merit pay would make us all like cave people fighting for a bone."*

Educators and policymakers frequently discuss ways to attract and retain high-quality teachers. One idea getting attention these days is to swap some of the benefits teachers enjoy later in their careers for more money in the early years. The survey finds teachers are protective of their pensions, and the vast majority of teachers overall do not like the idea of raising starting salaries in exchange for fewer retirement benefits. But many teachers are open to other new ways of attracting and keeping good teachers. Generally speaking, teachers appear to be considerably more interested in recruitment and retention strategies that would improve the flexibility and conditions of their work. For example, most support making it easier to leave and return to the profession without losing benefits. A suburban teacher from California wrote, "As a mom of two kids under five, I'd like to see it more feasible to take a few years off and be able to go back without retirement being so negatively affected." And an overwhelming majority supports giving teachers more time for class planning and preparation. While this measure would come with a large price tag for public schools, it is notable that the measure teachers are most likely to favor does not come with any monetary gain for individual teachers.

UNIONS AS PROTECTORS AND REFORMERS

Teachers unions play a powerful role in influencing the direction and success of district reforms aimed at improving teacher quality. Public school teachers expect unions to continue playing their traditional role: to bargain for benefits, safeguard jobs, and protect teachers from political machinations in their districts. But teachers also are open to their local union playing a role in improving teacher quality. While relatively few see the union in their own district as active in doing so, large numbers would support union efforts to mentor and train teachers, to negotiate new ways to evaluate teachers, and even to engage in high-stakes reform efforts such as guiding ineffective teachers out of the profession.

Unions Matter

Most teachers see the teachers union as vital to their profession. When asked how they think of teachers unions or associations, 54 percent of teachers responded that they are "absolutely essential." This is an increase of 8 percentage points from 46 percent in 2003. Another 31 percent see unions as "important but not essential," and just 11 percent as "something [they] could do without." Among union members, almost 2 out of 3 (65 percent) view them as absolutely essential. [Fig. 3–1]

Figure 3–1. Still Critical

Do you think of teachers unions or associations as: [Question 36]		
	Total 2003 (n=1,345)	Total 2007 (n=1,010)
Absolutely essential	46*	54*
Important but not essential	38*	31*
Something you could do without	12	11

*Statistically significant difference.

And most teachers strongly value the traditional protections that unions offer. Approximately three in four teachers (74 percent) agree that, "Without collective bargaining, the working conditions and salaries of teachers would be much worse." This has declined by 7 percentage points, from 81 percent who agreed in 2003. Not surprisingly, union members are far more likely than non-members to feel this way (87 percent vs. 50 percent).

Seventy-four percent agree that, "Teachers facing unfair charges from parents or students would have nowhere to turn without the union." Union members are more than twice as likely as their non-union colleagues to feel this way (85 percent vs. 39 percent).

Almost eight in 10 teachers (78 percent) agree that, "Without a union, teachers would be vulnerable to school politics or administrators who abuse their power." Again, union members are twice as likely to feel this way compared with non-members (91 percent vs. 45 percent).

Finally, most teachers do not think that union presence hinders the reputation of the profession. Just 21 percent of teachers agree that, "Teachers would have more prestige if collective bargaining and lifetime tenure were eliminated." Sixty percent of teachers overall *disagree* with this statement, as do 68 percent of union members and a smaller 44 percent of non-members.

On the Union Agenda

Public school teachers rely on their unions mainly for traditional functions. More than three quarters of teachers say that their local union "protects teachers through due process and grievance procedures," "regularly informs teachers about their benefits, rights, and responsibilities," and "effectively negotiates contracts, salary, and benefits on behalf of teachers." And of those teachers who report that their local union performs such traditional functions, most say it is doing an excellent or good job (approximately seven out of 10). While only 8 percent of teachers said they had filed a grievance against their district, the majority (73 percent) of these teachers reported that their only or most recent grievance ended in their favor. And 70 percent of these same teachers said the union did a good job representing them, while only 27 percent felt the union could have worked a lot harder.

Some unions, however, are moving outside of the traditional role and engaging in activities typically associated with a more vigorous school reform agenda. Fifty-five percent of teachers overall say the union in their district "negotiates to keep class size down in the district." Nearly half of teachers (46 percent) say that the local union "provides support and mentoring to new teachers." Forty-one percent say it "negotiates new ways to more meaningfully and effectively evaluate teachers" and that it "keeps teachers updated on new instructional methods and curriculum." Almost four in 10 (38 percent)

say their district's union "provides teachers with high-quality training and professional development," and one in three (33 percent) that it "expands the career ladder for teachers by negotiating new and differentiated roles and responsibilities."

Still, according to these survey results, most unions do not appear to be engaged in efforts to deal with ineffective teachers. Only 17 percent of teachers say that the union in their district "leads efforts to identify ineffective teachers and retrain them." Fifteen percent (for both) say that the union "guides ineffective teachers out of the profession" or "screens teachers who are new or transferring to ensure a good fit with the schools they're going to." [Fig. 3–2] Half of teachers (49 percent) agree that their union "sometimes fights to protect teachers who really should be out of the classroom." And nearly half (46 percent) say they "personally know a teacher in their building who is past the probationary period but who is clearly ineffective and shouldn't be in the classroom."

Figure 3–2. The Unions and Teacher Quality

Percent who say the union in their district currently does each item:

	Total (n=1,010)
Provide support and mentoring to new teachers [Question 58]	46
Negotiate new ways to more meaningfully and effectively evaluate teachers [Question 64]	41
Keep teachers updated on new instructional methods and curriculum [Question 55]	41
Provide teachers with high-quality training and professional development [Question 59]	38
Expand the career ladder for teachers by negotiating new and differentiated roles and responsibilities [Question 61]	33
Lead efforts to identify ineffective teachers and retrain them [Question 63]	17
Guide ineffective teachers out of the profession [Question 62]	15
Screen teachers who are new or transferring to ensure a good fit with the schools they're going to [Question 65]	15

Room to Grow

While teachers value unions for their traditional protections, sizeable numbers also seem open to the union as a player in reform. When forced to choose, more than half of teachers (52 percent) prefer that their union stick to traditional issues such as protecting teachers' salaries, benefits, and jobs. But nearly a third (32 percent) say that unions should increase their focus on things like teacher quality and student achievement (16 percent are unsure). [Fig. 3–3]

Figure 3–3. Remember Bread and Butter

Do you think that teachers unions should: [Question 39]

Among the teachers who say that the union or association in their district currently *does not* perform certain functions typically associated with a more vigorous school reform agenda, sizeable numbers would strongly favor their local union taking on such activities. For example, while 38 percent of teachers report that their local union doesn't provide support and mentoring to new teachers, the majority of these teachers indicate that the union *should* take on this responsibility (66 percent would favor the union doing so). And, while 39 percent of teachers report that their union currently doesn't negotiate new ways to more meaningfully and effectively evaluate teachers, the majority (72 percent) of these teachers would favor the union doing so.

The same pattern continues across other reforms. Among those who say the union in their district does not currently do so, approximately two out of three would favor their local union playing a role in guiding ineffective teachers out of the profession (66 percent), in expanding the career ladder for teachers (65 percent), and in identifying and retraining ineffective teachers (65 percent). Six in 10 support the union getting more involved in providing guidance on instructional and curriculum matters (61 percent) and also providing professional development opportunities (61 percent).

The one exception where there is less than a majority in favor is for the union to take a more active role in ensuring a good fit between teachers and schools; for this item, just under half (47 percent) say they are in favor. [Fig. 3–4]

Permission to Lead

Sizeable numbers of public school teachers indicate strong support for teachers unions to take the initiative on what many would consider to be controversial reforms.

More than six in 10 (63 percent) teachers in the overall sample say they would support the union or association

Figure 3–4. How Unions Can Improve Teaching

†Figures in parentheses represent totals.

Base: Teachers who say union currently does not do each item.

in their district taking the lead on ways to simplify the process for removing teachers who are past the probationary period and who are clearly ineffective and shouldn't be in the classroom. Just 16 percent would oppose it. Majorities of both union members (65 percent) and non-members (58 percent) are in favor of this idea.

Also, more than half of teachers overall (52 percent) say they would support the local union or association taking the lead on negotiating a way to add teacher performance as a consideration when deciding an individual teacher's salary; 40 percent would oppose it. [Fig. 3–5]

Figure 3–5. The Unions and Pay for Performance

Voices From the Field …

On Unions as Protectors:
"I would never give up my continuing contract rights," a 31-year veteran wrote. *"I have seen too many parents and administrators make unfounded accusations that could ruin a career."*

"Without our union, teachers are quite powerless," wrote a teacher from Hawaii.

"One of the reasons that I belong to the union, as ineffective as it may be … I belong because of the liability policy. If you're going to be a teacher, you need to have that liability. There are so many situations I couldn't even begin to name. If you coach, do an activity, or something in the classroom, I just believe that they have lawyers that are specially trained for the educational system, not just somebody who went to law school and can interpret law, but somebody that really knows educational law," said a Phoenix teacher.

"I have to say I just don't know what it would be like if we didn't have a union," said one New York City teacher. *"I'm losing faith in the union more and more all the time, but I don't know. … What would it be, if we didn't have one?"*

On Unions as Reformers:
Through the focus groups, it became evident that teachers fear losing the services and protections they have come to expect from unions if the unions were to take on more responsibilities. A Milwaukee area teacher was explicit: *"I would worry [that] my union couldn't handle taking on anything else. … Get me a contract. We haven't had a contract in 10 years. Then think about something else."*

Many teachers felt that it would be acceptable for their union to take on new things as long as old things keep getting done. *"I just think that there are other things that are more true to my feelings or my concerns. … If they were involved with [teacher quality/student achievement], I mean honestly, I would still probably have my focus or my concerns with my salary, the work day, those kinds of things,"* said a Phoenix teacher.

Cooperation and Conflict

A third (33 percent) of the union members surveyed for this study say the relationship between district leadership and the teachers union is mostly about "conflict and distrust." But the plurality (47 percent) characterizes the relationship as one of "cooperation and trust." [Fig. 3–6]

Summary and Analysis

The majority of public school teachers continue to view teachers unions as vital, and based on the survey results, membership continues to be strong, although participation levels vary. (See sidebar, The Union Way,

Figure 3–6. Conflict or Cooperation

Today in your district, how would you describe the relationship between the teachers union or association and the district leadership? Is it mostly about: [Question 40]

Base: Union member (n=671)

Page 11.) But loyalty to the union seems borne more of immediate practical concerns than a broader sense of unionism. "They're the policemen [sic] who just keep an eye on the laws and regulations," explained one teacher. And to a large extent, this may be how most teachers generally experience their union or association—as a necessary protector of their rights in an environment that often seems disconnected from, if not hostile to, their daily work lives.

As such, teachers tend to rely on their unions mainly for traditional bread-and-butter issues—securing money, benefits, and legal representation—and teachers report considerable satisfaction with their unions on these matters. Similar to other professionals, teachers worry about the increasing costs of health and dental insurance, about retirement—and they're counting on the union to protect those benefits. Teachers often talk about feeling extremely vulnerable to the powers that be—parents, principals, legislators. Partly for these reasons, and despite what teachers sometimes see as the unions' shortcomings, teachers continue to be tethered to the traditional role of unions.

Teachers do not appear to automatically associate their unions with efforts to improve teacher quality. In a New York City focus group, for example, teachers mentioned many recent instructional improvements such as smaller learning academies and curriculum compacting—but, rightly or wrongly, they did not attribute them to union initiatives. When asked specifically where the teachers union fit in, several teachers in the group identified as union-initiated a program that provides teacher coaching and a union-designed violence prevention workshop. But they did not intrinsically associate their unions with substantive initiatives until the moderator probed in this direction. As one New York City teacher said, "I never looked to the union for professional development. If I did [look to the union] it was a contractual question or something like that." Teachers also revealed confusion about the role unions should play in supporting innovation such as charter schools. (See sidebar, Charter Confusion, Page 12.)

The Union Way

Based on these survey results, the vast majority of public school teachers continue to value union membership, although most union members do not participate actively in their unions. In 2003, 83 percent of teachers reported that they were members of a teachers union or association; in 2007, the number remains virtually unchanged at 82 percent. But large majorities—approximately two out of three members—say they are not involved or engaged with the local union other than to receive mailings and notices (66 percent in 2003 and 69 percent in 2007).

There are disparities among union members on how well teachers unions reflect the views of most teachers—and how effective teachers unions are in general. Slightly more than half (51 percent) of the union members surveyed are of the opinion that when their union negotiates with district leadership, it "virtually always works for the best interests of its members and reflects their preferences," compared with just 18 percent that say the union "sometimes takes positions that are not in the best interests of its members or not aligned with what members want." Another 16 percent say it does both equally, and 14 percent are not sure. On three out of five initiatives regarding differentiated pay for teachers, union members are substantially less optimistic than their non-union counterparts.

While majorities of union members view the various levels of the union—building, district, state, national—as effective, a larger percentage points to the *district* level as effective compared with the others. Eighty-five percent of union members say the teachers union or association at the district level is effective. This is followed by 78 percent who say the union in their building is effective; 68 percent the state level; and 57 percent the national level.

Merit Pay by Union Membership

Percent who "strongly" or "somewhat" favor giving financial incentives to:

	Union Member (n=671)	Non-Member (n=165)
Teachers who work in tough neighborhoods with low-performing schools [Question 23]	79	79
Teachers who receive accreditation from the National Board for Professional Teaching Standards [Question 21]	63	67
Teachers who consistently receive outstanding evaluations by their principals [Question 19]	52*	71*
Teachers who specialize in hard-to-fill subjects such as science or mathematics [Question 22]	49*	61*
Teachers whose kids routinely score higher than similar students on standardized tests [Question 20]	28*	47*

*Statistically significant difference.

Charter Confusion

Public school teachers are about as likely to favor (42 percent) as they are to oppose (45 percent) the fundamental idea behind charter schools—schools that "operate under a charter or contract that frees them from many of the state regulations imposed on public schools and permits them to operate independently." In comparison, the public at large is somewhat more likely to favor charters, according to a recent survey that found 53 percent in favor and 34 percent in opposition.*

Despite the more than 4,000 charter schools that operate nationally and the intense debate that surrounds them, public school teachers appear to know little about charter schools.† Only 22 percent of teachers say they know "a great deal" or "quite a bit." Nearly four in 10 (39 percent) know either "very little" or "nothing at all." This is only slightly better than the public's knowledge level. A recent survey shows that 12 percent of registered voters, for example, know a lot about charter schools.‡

Teachers' unfamiliarity with charter schools is surprising given the footprint that charters now have in many communities. Findings from both the focus groups and the survey revealed confusion about charter schools and the role unions might play in managing or sponsoring them. The survey found that public school teachers are somewhat more likely to support (34 percent) than oppose (26 percent) having teachers unions themselves sponsor and manage charter schools, but the plurality (40 percent) are not sure—an unusually high percentage that indicates unsettled views. In New York City, where the teachers union is championing several charter school initiatives, only one teacher in the focus group seemed to have heard about it at all and announced that "the jury is still out." In Milwaukee, where charter schools are relatively prevalent, many teachers appeared to be uninformed and uninterested in the topic. As union leaders think about organizing charter schools or sponsoring schools of their own, they have considerable work to do to educate their membership.

Don't Know Much About Charter Schools?

How much do you know about charter schools? [Question 82]

*38th Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes Toward the Public Schools, 2006.

†According to the Center for Education Reform, the exact number is 4,147 as of 2007 (http://www.edreform.com/_upload/CER_charter_numbers.pdf).

‡Glover Park Group, 2005.

According to teachers, unions do not appear to be particularly active on the teacher quality front, although many teachers indicate support for this type of union activity. Thus, teachers unions have a lot of room to expand the role they play in improving teacher quality. Initiatives such as mentoring new teachers or serving as a resource on curriculum and teaching methods are low-hanging fruit—fairly easy to implement and relatively non-controversial.

Overall, teachers are fairly receptive to expanding the union role in reform, especially when it comes to improving the state of their craft. The findings strongly suggest teachers would back the union in their district if it were to take on such things as high-quality professional training or if it tried to expand the career ladder for teachers. And teachers seem willing to go even further. They'd want to see their union working toward new ways to effectively evaluate teachers—and even to guide ineffective colleagues out.

Still, this is not a tame agenda for unions, and they do not pursue it without some risk. For the unions to take on all of these things at the same time might be overwhelming and may raise questions about their ability to deliver on the traditional issues that teachers say matter most. Moreover, it's one thing for teachers to voice support for an initiative or idea in a survey, quite another to do so in real life when there are high stakes attached.

NEWCOMERS AND VETERANS

Teachers with fewer than five years of experience (newcomers) and those with more than 20 years (veterans) agree on many issues. They both, for instance, value unions and the more traditional services they provide. But newer teachers and veteran teachers have substantially different attitudes toward differentiated pay as well as other aspects of their profession. Teachers also differ in opinions according to the regions in which they live—South, Northeast, West, Midwest. (See sidebar, Southern Comfort, Page 13.)

Shared Values

Both veteran teachers and newcomers value unions, especially their role in safeguarding teachers' jobs. Veterans are more likely than newcomers to say the

Southern Comfort

The survey findings strongly suggest that public school teachers in the South are more willing to push for reform than their peers in the Northeast, where union tradition remains stronger. For example, teachers working in southern states are more likely to favor pay for performance measured by student standardized test scores (44 percent, compared with 21 percent Northeast, 30 percent West, and 30 percent Midwest).

Also, there is a clear pattern of stronger support for the value-added approach for measuring teacher effectiveness among those in the South (55 percent, compared with 44 percent Northeast, 50 percent West, and 44 percent Midwest).

These regional differences may be more a function of union penetration than anything else, as teachers are far more likely to be union members in the Northeast and far less likely in the South. On virtually all of the questions in the survey pertaining to union activities, teachers in the Northeast are more likely than those in the South to say the union in their district currently takes part, that it is doing a good job, or that they would favor the union taking on that responsibility. In a nutshell, pro-union sentiment is prevalent among teachers in the Northeast, lacking among teachers in the South, and tends to fall somewhere in between for teachers in the West and Midwest.

Merit Pay by Region

Percent who "strongly" or "somewhat" favor giving financial incentives to:

	Northeast (n=163)	South (n=342)	West (n=200)	Midwest (n=266)
Teachers who work in tough neighborhoods with low-performing schools [Question 23]	75*	80	86*	78
Teachers who receive accreditation from the National Board for Professional Teaching Standards [Question 21]	59*	69*	65	61
Teachers who consistently receive outstanding evaluations by their principals [Question 19]	48*	65*	56	54
Teachers who specialize in hard-to-fill subjects such as science or mathematics [Question 22]	42*	60*	57	49
Teachers whose kids routinely score higher than similar students on standardized tests [Question 20]	21*	44*	30	30

*Statistically significant difference.

What Unions Do for Me

Percent of teachers who:

	Northeast (n=163)	South (n=342)	West (n=200)	Midwest (n=266)
Agree that without a union, teachers would be vulnerable to school politics or administrators who abuse their power [Question 49]	93	65*	83	82
Agree that without collective bargaining, the working conditions and salaries of teachers would be much worse [Question 48]	88	60*	80	81
Agree that the union regularly provides information and opportunities to help them be a better teacher [Question 44]	52*	37	38	41
Say that being a union member provides feelings of pride and solidarity, in addition to the practical benefits [Question 77]	44*	25	29	29
Believe that the type of school that would be better for children is one where work rules and school duties affecting teachers are defined by contract [Question 66]	35*	15	21	21

*Statistically significant difference.

teachers union is "absolutely essential" (60 percent compared with 51 percent). And, notably, newer teachers are considerably more likely to say the union is absolutely essential than they were four years ago (51 percent in 2007 compared to 30 percent in 2003). [Fig. 4–1]

Figure 4–1. The Continued Importance of Unions

Do you think of teachers unions or associations as: [Question 36]

	Newcomer		Veteran	
	2003 (n=211)	2007 (n=110)	2003 (n=484)	2007 (n=363)
Absolutely essential	30*	51*	57	60
Important but not essential	49*	32*	30	27
Something you could do without	13	11	11	11

*Statistically significant difference.

Note: Newcomer=less than five years; Veteran=more than 20 years.

Although a quarter of newcomers think unions lean toward tending to the needs of veteran teachers (26 percent), large numbers of both groups see no patterns of favoritism. Forty-six percent of newer teachers and 68 percent of veteran teachers believe that the union in their district focuses about equally on both groups. Few think the needs of new teachers get most of the attention (4 percent of newcomers and 10 percent of veterans). Fully one in four (25 percent) newer teachers is not sure, compared with 13 percent of veterans. [Fig. 4–2]

More Positive About Union Protections

It is not surprising that teachers with more than 20 years of experience would be more active in the teachers union or to perceive the union as acting in its members' best

Figure 4–2. Favoritism?

*Statistically significant difference.

Note: Newcomer=less than five years; Veteran=more than 20 years.

interests. For example, among those who say the union in their district protects teachers through due process and grievance procedures, veterans are more likely to think the union is doing an excellent or good job (78 percent vs. 62 percent of newcomers). And veteran teachers are also more likely than their newer counterparts to agree that "without a union, teachers would be vulnerable to school politics or administrators who abuse their power" (84 percent vs. 72 percent); that "without collective bargaining, the working conditions and salaries of teachers would be much worse" (80 percent vs. 69 percent); that "the union regularly provides information and opportunities to help them be a better teacher" (48 percent vs. 39 percent); and that "being a union member provides feelings of pride and solidarity, in addition to the practical benefits" (39 percent vs. 23 percent). [Fig. 4–3]

Figure 4–3. On My Side

*Statistically significant difference.

Note: Newcomer=less than five years; Veteran=more than 20 years.

More Open to Reforms

Compared with veterans, newer teachers are more supportive of a range of reforms that would reward existing teachers for superior performance or recruit new high-quality candidates to the profession. On each of five proposals posed in the survey about giving financial incentives to teachers, newcomers are more likely than veterans to be positive. [Fig. 4–4]

Figure 4–4. Favoring Financial Incentives

*Statistically significant difference.

Note: Newcomer=less than five years; Veteran=more than 20 years.

Most newcomers (65 percent) also support the union taking "the lead on negotiating a way to add teacher performance as a consideration when deciding an individual teacher's salary." Only 45 percent of veteran teachers are in favor.

While 58 percent of newcomers believe there are outstanding teachers in their school "who deserve to be especially rewarded because they do a stellar job," only 39 percent of veterans agree. Newcomers are far more likely than veterans to think it is an excellent or good idea to "pay signing bonuses to attract new talent" (65 percent vs. 37 percent). Newcomers are also more likely to react positively toward "offering new teachers substantially higher starting salaries in exchange for smaller pensions when they retire (30 percent vs. 20 percent).

Finally, 58 percent of newcomers think it is an excellent or good idea to "actively recruit and train second-

career candidates from business and other sectors of the economy," compared with 50 percent of veterans. [Fig. 4–5]

Figure 4–5. Improving the Profession

*Statistically significant difference.

Note: Newcomer=less than five years; Veteran=more than 20 years.

Summary and Analysis

In the public debate it is generally assumed that novice teachers are more skeptical of teachers unions and more open to change than veteran teachers. This is true in some ways, but not in all—the picture is far more complicated.

Veterans do express stronger positive sentiments than newcomers about the teachers union and the protections it offers. With more years in the system, veteran teachers are more likely to have witnessed the union defend their colleagues or themselves against what they perceive to be onerous or nonsensical work rules. Thus, they may be more aware of the value that unions offer. Likewise, over time in a career people tend to become more averse to change and risk, and veterans may see the union as a helpful bulwark against those things.

Yet, despite general assumptions, newcomers continue to view teachers unions as absolutely essential. In fact, over the past four years, an increasingly larger percentage of newer teachers say they view teachers unions as absolutely essential (51 percent in 2007 vs. 30 percent in 2003); among veterans there was virtually no change (60 percent in 2007 and 57 percent in 2003). And newcomers are still attached to the union's traditional functions. A majority of newcomers say the union should "mostly stick to traditional union issues" rather than "put more focus than they currently do" on reform-centered activities (59 percent vs. 29 percent).

The enduring appeal of unions to newer teachers could be the result of a number of things. Teachers may appreciate having union backing in a more contentious No Child Left Behind Act era, where the public schools—and teachers themselves—are under greater scrutiny than ever before. It may also be that as newer teachers perceive union power as on the decline, they may be more inclined to nostalgically reflect about its usefulness. Or it may be that today's newer teachers are responding to broader economic and workplace changes. A Pew Research Center survey in 2006 showed that workers are more likely to worry that their employers are less loyal to them and that their jobs provide fewer benefits. In any case, newcomers and veterans alike may express support for teachers unions—warts and all—because they truly believe they need the protection they offer for things like salary and benefits.

But newcomers also have more doubts than veterans about how helpful unions really are. They are less likely to think the union offers protections from abusive administrators or safeguards the working conditions and salaries for teachers. They are less likely to think—and perhaps in a better position to know—that the union helps them be better teachers. Compared to veterans, newcomers are less likely to think the union always acts in the best interest of its members, which may partially explain why newcomers are also less inclined to feel pride or solidarity in regards to union membership or to be active in their local union.

Most importantly, newcomers are considerably more open to some reform-minded initiatives. They are amenable even to the more controversial proposals, that is, the ones involving the use of student achievement to determine teacher pay. There are also significant distinctions between the two groups on the overall topic of differentiated pay for teachers.

CONCLUSION

Various parties in the education debate often claim to know what teachers want or to speak on their behalf. Public school teachers' views, however, are hardly unanimous or monolithic; they are nuanced and sometimes even contradictory, which speaks to the complexity of the issues and the fact that reasonable people can disagree. The results of this survey clearly show that reformers, school districts, and teachers unions all have their work cut out for them if they truly want to lay claim to the support of the nation's teachers.

This survey points to several important takeaways. First, before the reform conversation can even get started, school district management must meet its core obligations to create a well-functioning workplace for teachers. For their part, the unions must take on, in a meaningful way, some of the chronic problems that damage their public brand, frustrate teachers, and have an adverse impact on students. Labor and management must find ways to work together and advance a reform agenda. Ultimately, their fortunes are intertwined.

Policymakers and policy advocates must become more effective in how they communicate with teachers and explain reform ideas, particularly those ideas that are more controversial. White papers and reports are a thin reed and limited vehicle for sharing information in an environment where multiple institutions are seeking to communicate with teachers and win them over. Reform ideas must be communicated in multiple ways and to multiple audiences. Today, with the increasing prevalence of electronic forms of communication, this is more possible than ever before.

That the loyalty of K–12 public school teachers is up for grabs is ultimately an opportunity for education advocates, teachers unions, and policymakers but most importantly for the nation's current and future teachers. Research shows that teachers are the most important in-school factor affecting student achievement, but neither practice (in most schools and school districts) nor policy (local, state, or federal) is yet aligned with that finding. A vigorous debate about how to transform schools and teaching to meet today's challenges and create a profession that people seek to be part of, rather than one where they feel they need protection from unfair and capricious practices, is a vital one. The findings presented here, while not the last word, offer guideposts for that conversation.

Appendix A. Methodology

Waiting to Be Won Over **is based on a nationally representative random sample of 1,010 K–12 public school teachers conducted in fall 2007. The margin of error for the overall sample is plus or minus 3 percentage points. The survey was preceded by six focus groups.**

The Survey

The sample was randomly drawn from a comprehensive database of current K–12 teachers maintained by Market Data Retrieval, a subsidiary of Dun & Bradstreet. A multi-mode approach that included both mail and online versions of the survey was used.

The first mailing, which included a questionnaire and a cover letter, was sent via first-class mail on September 25, 2007, to 7,200 randomly selected K–12 public school teachers in the United States. A reminder postcard was sent on October 3, 2007. A second mailing of the questionnaire with instructions to those who had not yet participated was sent on October 16, 2007. Each mailing of the questionnaire included a prepaid business reply envelope. Teachers for whom an e-mail address was available—about 50 percent of the sample—were sent e-mails with an embedded URL linking them to the online version. A total of 139 surveys were completed online.

The margin of error for the results from the overall sample of 1,010 is plus or minus 3 percentage points. It is higher when comparing percentages across subgroups. Subgroup differences reported in this study are statistically significant at the 95 percent confidence level. Teachers in urban districts were oversampled to insure the survey netted a sufficient number (1,200 of the original 7,200 were part of the urban oversample). The results are weighted to reflect the actual distribution of urban teachers in the teacher population.

The overall response rate for the survey is 14 percent. As with all surveys, the risk of non-response is that the pool of survey respondents could differ from the true population of teachers, decreasing the ability to draw inferences from the data. A comparison of the demographic profile of respondents to that of the overall population of teachers shows they are very similar when it comes to such key variables as race and ethnicity, urbanicity, region, and sex (see Population vs. Sample Comparison). Results can also be affected by non-sampling sources of bias, such as question wording. Steps were taken to minimize these, including extensive pre-testing of the survey instrument with focus group participants and six one-on-one telephone interviews with current K–12 teachers.

The questionnaire (see Appendix B) was designed by the FDR Group and Education Sector; the two organizations are responsible for all interpretation and analysis contained within this report. FDR Group (Farkas Duffett Research Group) is a nonpartisan public opinion research firm specializing in surveys, focus groups, and program evaluations. The survey was fielded and tabulated by Robinson and Muenster Associates, Inc., of Sioux Falls, South Dakota.

The Focus Groups

To help develop the questionnaire, six focus groups with K–12 public school teachers were conducted, with each group having 10–12 participants. The groups were conducted in five sites selected for geographic and regional representation: Milwaukee (one group with teachers working in the city, another with teachers working in the suburbs), New York City (teachers working in city only), Chicago (mix of city and suburban), Atlanta (mix of city and suburban), and Phoenix (mix of city and suburban). Participants were recruited to FDR Group specifications to ensure a proper demographic mix. These discussions were crucial to developing the wording of the survey questions and to understanding why teachers feel as they do. Quotes in this report are drawn from the focus groups and from comments survey respondents wrote on their questionnaires in response to open-ended questions. All focus groups were moderated by the FDR Group.

Population vs. Sample Comparison (by percent)

	Population	Sample (n=1,010)
Race/Ethnicity		
White	83	88
Black	8	5
Hispanic	6	4
Asian/Pacific	2	1
Native American/Other	2	2
Sex		
Male	25	21
Female	75	79
School Type		
Elementary	52	51
Middle	20	21
High	23	27
Something else	5	1
Urbanicity		
Urban	31	29
Suburban	38	42
Rural/small town	31	29
School Enrollment		
<300	11	10
300–499	23	26
500–999	45	36
1,000 or more	22	28
Region		
Northeast	18	17
Midwest	24	28
South	39	35
West	19	20

Sources: US Department of Education, Institute of Education Science, National Center for Education Statistics, Digest of Education Statistics, 2006; NCES School and Staffing Survey, 2003–2004.

Appendix B. National Survey of Public School Teachers

This survey is based on a national random sample of 1,010 K–12 public school teachers. It was conducted by mail and online in fall 2007. The margin of error is plus or minus 3 percentage points. Numbers may not add up to 100 percent due to rounding. An asterisk (*) indicates less than one percent; a dash (-) indicates zero.

1. Are you:
79 Female
21 Male

2. Which best describes your current teaching position:
98 A full-time teacher in a traditional public school
1 A full-time teacher in a charter school
1 Something else

3. Do you currently teach at:
51 Elementary school
21 Middle school or Junior high school
27 High School
1 Something else

4. For how many years have you been a PUBLIC school teacher?
11 1–4 years
20 5–9 years
33 10–20 years
37 21 years or more

5. What subject or subjects do you primarily teach? [Check all that apply.]
31 All subjects
22 English and/or Reading
19 Mathematics
16 Social Studies or Social Sciences
14 Science
9 Physical Education or Health
8 Art, Music or Fine Arts
7 Special Education/Gifted/ESL
3 Computer Science
2 Foreign Language
1 Business
7 Something else

How much do you agree or disagree with the following statements about teachers and the public schools? [Questions 6–11]

6. Although they are on the front lines, teachers are rarely consulted about what happens in their schools
35 Strongly Agree
45 Somewhat Agree
14 Somewhat Disagree
6 Strongly Disagree
* Not Sure

7. Teachers are required to do too much paperwork and documentation about what goes on in their classrooms
49 Strongly Agree
37 Somewhat Agree
11 Somewhat Disagree
3 Strongly Disagree
* Not Sure

8. Too many veteran teachers who are burned out stay because they do not want to walk away from the benefits and service time they have accrued
40 Strongly Agree
36 Somewhat Agree
13 Somewhat Disagree
7 Strongly Disagree
5 Not Sure

9. Too much negative press coverage about the public schools discourages talented, well-educated people from pursuing teaching as a career
42 Strongly Agree
39 Somewhat Agree
13 Somewhat Disagree
3 Strongly Disagree
3 Not Sure

10. When individual schools fail it's usually because they have ineffective principals at the helm
7 Strongly Agree
33 Somewhat Agree
35 Somewhat Disagree
21 Strongly Disagree
5 Not Sure

11. All the paperwork and legal and contractual restrictions make it difficult for principals to get things done
15 Strongly Agree
44 Somewhat Agree
20 Somewhat Disagree
8 Strongly Disagree
13 Not Sure

How would you rate each of the following ideas for attracting and retaining high-quality teachers to the teaching profession? [Questions 12–18]

12. Actively recruit and train second-career candidates from other fields and sectors of the economy
17 Excellent
37 Good
30 Fair
13 Poor
4 Not Sure

13. Ease state certification requirements while intensifying in-class supervision, observation, and mentoring
14 Excellent
29 Good
26 Fair
29 Poor
3 Not Sure

14. Give teachers more time during the school day for class preparation and planning
57 Excellent
28 Good
11 Fair
4 Poor
* Not Sure

15. Make it far easier to leave and return to teaching without losing retirement benefits
44 Excellent
34 Good
14 Fair
6 Poor
3 Not Sure

16. Offer new teachers substantially higher starting salaries in exchange for smaller pensions when they retire

7 Excellent
15 Good
23 Fair
48 Poor
7 Not Sure

17. Pay signing bonuses to attract new talent

15 Excellent
27 Good
26 Fair
27 Poor
4 Not Sure

18. Make it easier to earn and take sabbatical leave for teachers working in really challenging schools

30 Excellent
40 Good
18 Fair
6 Poor
6 Not Sure

How much would you favor or oppose giving financial incentives to each of the following: [Questions 19–23]

19. Teachers who consistently receive outstanding evaluations by their principals

24 Strongly Favor
34 Somewhat Favor
18 Somewhat Oppose
21 Strongly Oppose
3 Not Sure

20. Teachers whose kids routinely score higher than similar students on standardized tests

11 Strongly Favor
23 Somewhat Favor
25 Somewhat Oppose
39 Strongly Oppose
3 Not Sure

21. Teachers who receive accreditation from the National Board for Professional Teaching Standards

25 Strongly Favor
40 Somewhat Favor
16 Somewhat Oppose
15 Strongly Oppose
4 Not Sure

22. Teachers who specialize in hard-to-fill subjects such as science or mathematics

17 Strongly Favor
37 Somewhat Favor
23 Somewhat Oppose
20 Strongly Oppose
4 Not Sure

23. Teachers who work in tough neighborhoods with low-performing schools

34 Strongly Favor
46 Somewhat Favor
11 Somewhat Oppose
7 Strongly Oppose
3 Not Sure

24. Suppose that in your district the students of some teachers make more academic progress—in terms of improved reading levels, teacher evaluations, and classroom tests—when compared to similar students taught by other teachers. How much would you favor or oppose financially rewarding those teachers?

10 Strongly Favor
34 Somewhat Favor
22 Somewhat Oppose
29 Strongly Oppose
5 Not Sure

25. At your school, do you think there are outstanding teachers who deserve to be especially rewarded because they do a stellar job?

48 Yes
5 No
40 There are outstanding teachers, but I don't think they should be especially rewarded
7 Not Sure

26. In what ways, if any, do school officials at your school or district reward outstanding teachers? [Check all that apply.]

5 Financial bonus
16 Informal recognition (for example, better treatment or perks)
29 Official recognition (for example, formal commendation or note to file)
10 Token gift
49 They do not reward outstanding teachers; the reward is solely intrinsic
10 Not Sure

27. Some suggest that the best way to measure teacher effectiveness is to assess students' skills and knowledge when they first come to a teacher and to measure them again when students leave to see what progress was made. How would you rate this as a way of measuring teacher effectiveness?

15 Excellent
34 Good
29 Fair
20 Poor
2 Not Sure

28. Thinking of your own experience being evaluated as a teacher, which statement would come closest to describing your most recent formal evaluation?

26 It was useful and effective in terms of helping you be a better teacher
32 It was well-intentioned but not particularly helpful to your teaching practice
41 It was just a formality
2 Not Sure

The next few questions are about tenure. Although "tenure" policies vary from state to state, for the purposes of this survey, please think of a tenured teacher as one who has been awarded job protections and due process rights after successfully completing a probationary period, typically 2 to 4 years in length.

29. Are you currently a tenured teacher, or not?

64 Yes, a tenured teacher
15 Yes, it's not called tenure, but I have job protections and due process rights
14 No, not a tenured teacher
6 No, there is no tenure at my school
2 Not Sure

30. In general, when you hear that a teacher at your school has been awarded tenure, which of these two thoughts would be more likely to cross your mind?

23 That the teacher has proven to be very good at what s/he does
69 That it's just a formality—it has very little to do with whether a teacher is good or not
8 Not Sure

31. To what extent would you support or oppose strengthening the formal evaluation of probationary teachers so that they will get tenure only after they've proven to be very good at what they do?

38 Strongly Support
41 Somewhat Support
10 Somewhat Oppose
3 Strongly Oppose
8 Not Sure

32. And when it comes to tenured teachers, how often do you think they should be formally evaluated?

32 Yearly
26 At least every 2 years
22 Every 3–4 years
13 Every 5 years
1 Never
3 Something else
3 Not Sure

33. If you had the choice, would you personally be willing to trade tenure for a pay increase (e.g., $5,000 per year), or would the pay increase have to be a lot higher, or would you rather hold on to tenure? *[Base: Tenured Teachers]*

25 Would trade tenure for a pay increase
29 Would have to be a lot higher
29 Would rather hold on to tenure
17 Not Sure

34. If you had the choice, would you personally be willing to trade tenure for more autonomy and control over decisions affecting your school, would it have to be a lot more autonomy and control, or would you rather hold on to tenure? *[Base: Tenured Teachers]*

10 Would trade tenure for more autonomy and control
18 Would have to be a lot more autonomy and control
53 Would rather hold on to tenure
20 Not Sure

35. Check the statement that best describes your current status:

68 I am a member of a teachers union or association that engages in collective bargaining
15 I am a member of a professional association that provides such things as liability insurance, but not collective bargaining
16 I am not a member of a teachers union or association
1 There is no teachers union or association to join in my district
1 Not Sure

Whether or not you are currently a member of a union or association, or whether collective bargaining exists in your district, please answer the remaining questions to the best of your knowledge. As a public school teacher, your opinion counts. Remember, if you feel an item is not applicable to you, please skip it and move on to the next one.

36. Do you think of teachers unions or associations as:

54 Absolutely essential
31 Important but not essential
11 Something you could do without
4 Not Sure

37. In many states, the teaching profession is unionized, and salary, benefits and work rules are determined by collective bargaining. When you chose to become a teacher, did this make the profession:

13 More appealing to you
5 Less appealing
79 Was not a consideration
3 Not Sure

38. Similarly, teaching is sometimes perceived as a profession with considerable job protection, one where it is rare to lose your job. When you chose to become a teacher, did this make the profession:

14 More appealing to you
1 Less appealing
84 Was not a consideration
2 Not Sure

39. Generally speaking, do you think that teachers unions or associations should:

32 Put more focus than they currently do on issues such as improving teacher quality and student achievement
52 Mostly stick to traditional union issues such as protecting teachers' salaries, benefits, and jobs
16 Not Sure

40. Today in your district, how would you describe the relationship between the teachers union or association and the district leadership? Is it mostly about:

28 Conflict and distrust
44 Cooperation and trust
4 There is no union or association
24 Not Sure

How much do you agree or disagree with the following statements? [Questions 41–49]

41. Teachers would have more prestige if collective bargaining and lifetime tenure were eliminated

4 Strongly Agree
17 Somewhat Agree
23 Somewhat Disagree
38 Strongly Disagree
19 Not Sure

42. Despite having the strength of their unions behind them, rank-and-file teachers usually have very little control over what goes on in their own schools

32 Strongly Agree
42 Somewhat Agree
16 Somewhat Disagree
4 Strongly Disagree
7 Not Sure

43. The union charges far higher dues than are warranted by what it does for teachers

21 Strongly Agree
33 Somewhat Agree
19 Somewhat Disagree
15 Strongly Disagree
12 Not Sure

44. The union regularly provides information and opportunities to help me be a better teacher

9 Strongly Agree
32 Somewhat Agree
26 Somewhat Disagree
21 Strongly Disagree
13 Not Sure

45. The union sometimes fights to protect teachers who really should be out of the classroom

14 Strongly Agree
35 Somewhat Agree
18 Somewhat Disagree
10 Strongly Disagree
24 Not Sure

46. Teachers facing unfair charges from parents or students would have nowhere to turn without the union

41 Strongly Agree
34 Somewhat Agree
12 Somewhat Disagree
5 Strongly Disagree
9 Not Sure

47. New teachers tend to place less value on the union

17 Strongly Agree
42 Somewhat Agree
10 Somewhat Disagree
5 Strongly Disagree
26 Not Sure

48. Without collective bargaining, the working conditions and salaries of teachers would be much worse

44 Strongly Agree
31 Somewhat Agree
7 Somewhat Disagree
4 Strongly Disagree
15 Not Sure

49. Without a union, teachers would be vulnerable to school politics or administrators who abuse their power

47 Strongly Agree
30 Somewhat Agree
9 Somewhat Disagree
4 Strongly Disagree
10 Not Sure

Overall, how effective would you say the teachers union or association is at the following levels: [Questions 50–53]

50. At the building where you work

22 Strongly Effective
42 Somewhat Effective
18 Not Too Effective
10 Not Effective At All
9 Not Sure

51. At the district level

25 Strongly Effective
48 Somewhat Effective
12 Not Too Effective
6 Not Effective At All
10 Not Sure

52. At the state level

18 Strongly Effective
44 Somewhat Effective
15 Not Too Effective
4 Not Effective At All
19 Not Sure

53. At the national level

13 Strongly Effective
39 Somewhat Effective
18 Not Too Effective
6 Not Effective At All
24 Not Sure

Here are some functions that teachers unions or associations may or may not perform. For each, please indicate whether the union or association in your district currently does it or not. Then, answer the corresponding "IF YES" or "IF NO" questions. [Questions 54–65]

54A. Effectively negotiate contracts, salary, and benefits on behalf of teachers—does the union or association in your district currently do this?

76 Yes
13 No
11 Not Sure/No Answer

54B. IF YES: How good a job is it doing?

26 Excellent
42 Good
25 Fair
6 Poor
2 Not Sure

54C. IF NO: Would you favor or oppose the union taking on this function?

29 Strongly Favor
28 Somewhat Favor
12 Somewhat Oppose
13 Strongly Oppose
19 Not Sure

55A. Keep teachers updated on new instructional methods and curriculum—does the union or association in your district currently do this?

41 Yes
46 No
13 Not Sure/No Answer

55B. IF YES: How good a job is it doing?

13 Excellent
39 Good
33 Fair
10 Poor
5 Not Sure

55C. IF NO: Would you favor or oppose the union taking on this function?

24 Strongly Favor
37 Somewhat Favor
14 Somewhat Oppose
13 Strongly Oppose
12 Not Sure

56A. Negotiate to keep class size down in the district—does the union or association in your district currently do this?

55 Yes
30 No
14 Not Sure/No Answer

56B. IF YES: How good a job is it doing?

13 Excellent
36 Good
28 Fair
18 Poor
5 Not Sure

56C. IF NO: Would you favor or oppose the union taking on this function?

56 Strongly Favor
27 Somewhat Favor
5 Somewhat Oppose
4 Strongly Oppose
9 Not Sure

57A. Protect teachers through due process and grievance procedures—does the union or association in your district currently do this?

84 Yes
4 No
12 Not Sure/No Answer

57B. IF YES: How good a job is it doing?

30 Excellent
40 Good
19 Fair
3 Poor
9 Not Sure

57C. IF NO: Would you favor or oppose the union taking on this function?

33 Strongly Favor
29 Somewhat Favor
4 Somewhat Oppose
11 Strongly Oppose
23 Not Sure

58A. Provide support and mentoring to new teachers—does the union or association in your district currently do this?

46 Yes
38 No
16 Not Sure/No Answer

58B. IF YES: How good a job is it doing?

18 Excellent
39 Good
28 Fair
7 Poor
8 Not Sure

58C. IF NO: Would you favor or oppose the union taking on this function?

33 Strongly Favor
33 Somewhat Favor
12 Somewhat Oppose
9 Strongly Oppose
13 Not Sure

59A. Provide teachers with high-quality training and professional development—does the union or association in your district currently do this?

38 Yes
46 No
16 Not Sure/No Answer

59B. IF YES: How good a job is it doing?

14 Excellent
41 Good
28 Fair
10 Poor
7 Not Sure

59C. IF NO: Would you favor or oppose the union taking on this function?

27 Strongly Favor
34 Somewhat Favor
15 Somewhat Oppose
11 Strongly Oppose
14 Not Sure

60A. Regularly inform teachers about their benefits, rights, and responsibilities—does the union or association in your district currently do this?

79 Yes
10 No
11 Not Sure/No Answer

60B. IF YES: How good a job is it doing?

27 Excellent
43 Good
25 Fair
4 Poor
1 Not Sure

60C. IF NO: Would you favor or oppose the union taking on this function?

38 Strongly Favor
35 Somewhat Favor
4 Somewhat Oppose
8 Strongly Oppose
15 Not Sure

61A. Expand the career ladder for teachers by negotiating new and differentiated roles and responsibilities—does the union or association in your district currently do this?

33 Yes
44 No
23 Not Sure/No Answer

61B. IF YES: How good a job is it doing?

12 Excellent
31 Good
30 Fair
14 Poor
14 Not Sure

61C. IF NO: Would you favor or oppose the union taking on this function?

27 Strongly Favor
38 Somewhat Favor
8 Somewhat Oppose
6 Strongly Oppose
22 Not Sure

62A. Guide ineffective teachers out of the profession—does the union or association in your district currently do this?

15 Yes
61 No
24 Not Sure/No Answer

62B. IF YES: How good a job is it doing?

5 Excellent
13 Good
21 Fair
22 Poor
39 Not Sure

62C. IF NO: Would you favor or oppose the union taking on this function?

36 Strongly Favor
30 Somewhat Favor
9 Somewhat Oppose
9 Strongly Oppose
16 Not Sure

63A. Lead efforts to identify ineffective teachers and retrain them—does the union or association in your district currently do this?

17 Yes
60 No
24 Not Sure/No Answer

63B. IF YES: How good a job is it doing?

6 Excellent
20 Good
22 Fair
18 Poor
35 Not Sure

63C. IF NO: Would you favor or oppose the union taking on this function?

34 Strongly Favor
31 Somewhat Favor
10 Somewhat Oppose
11 Strongly Oppose
14 Not Sure

64A. Negotiate new ways to more meaningfully and effectively evaluate teachers—does the union or association in your district currently do this?

41 Yes
39 No
20 Not Sure/No Answer

64B. IF YES: How good a job is it doing?

11 Excellent
34 Good
34 Fair
11 Poor
10 Not Sure

64C. IF NO: Would you favor or oppose the union taking on this function?

36 Strongly Favor
36 Somewhat Favor
8 Somewhat Oppose
8 Strongly Oppose
12 Not Sure

65A. Screen teachers who are new or transferring to ensure they are a good fit with the schools they're going to—does the union or association in your district currently do this?

15 Yes
66 No
19 Not Sure/No Answer

65B. IF YES: How good a job is it doing?

14 Excellent
18 Good
23 Fair
16 Poor
28 Not Sure

65C. IF NO: Would you favor or oppose the union taking on this function?

20 Strongly Favor
27 Somewhat Favor
19 Somewhat Oppose
17 Strongly Oppose
17 Not Sure

66. On the whole, which type of school do you think would be better for students?

21 One where work rules and school duties affecting teachers are defined by contract
68 One where principals and teachers have more control and flexibility over these matters
2 Something else
1 Both
8 Not Sure

67. In some districts, the process for removing teachers who are clearly ineffective and shouldn't be in the classroom—but who are past the probationary period—is very difficult and time-consuming. Is this the case in your district, or not?

55 Yes 13 No 32 Not Sure

68. Do you personally know a teacher in your building who is past the probationary period but who is clearly ineffective and shouldn't be in the classroom, or not?

46 Yes 42 No 12 Not Sure

69. Assume that teachers would keep some due process protection against unfair practices by administrators. If the union or association in your district were to take the lead on ways to simplify the process for removing such teachers, how much would you support or oppose the effort?

22 Strongly Support
41 Somewhat Support
10 Somewhat Oppose
7 Strongly Oppose
2 Union already does this
19 Not Sure

70. Which of these do you think is the most likely course of action a principal in your district would take if faced with a persistently ineffective teacher who was already past the probationary period?

14 Do nothing
18 Initiate formal proceedings to remove the teacher from the district's employ
26 Make a serious effort to retrain the teacher
13 Quietly encourage the teacher to leave
14 Transfer the teacher to another school in the district
15 Not Sure

71. Typically, teachers get salary increases according to a strictly defined schedule mostly driven by their years of service and the credits they attain. Is this mostly how it works in your district, or not?

97 Yes 2 No 1 Not Sure

72. Assume that years of service and number of credits would still be taken into account. If the union or association in your district were to take the lead on negotiating a way to add teacher performance as a consideration when deciding an individual teacher's salary, how much would you support or oppose the effort?

18 Strongly Support
34 Somewhat Support
18 Somewhat Oppose
23 Strongly Oppose
1 Union already does this
8 Not Sure

73. Some school districts have a system for matching teachers with schools where any teacher, regardless of seniority, has an equal opportunity to fill a vacancy. It basically comes down to whether the teacher wants to work in the school and whether the school wants the teacher. Is this mostly how it works in your district, or not?

44 Yes 38 No 19 Not Sure

74. If the union or association in your district was trying to move in this direction, how much would you support or oppose the effort?

17 Strongly Support
31 Somewhat Support
11 Somewhat Oppose
9 Strongly Oppose
4 Union already does this
20 Not Sure

75. Overall, when the union or association in your district negotiates with district leadership, does it:

43 Virtually always work for the best interests of its members and reflect their preferences
17 Sometimes take positions that are not in the best interests of its members or not aligned with what members want
15 Both equally
26 Not Sure

76. Would you say that the teachers union or association in your district leans more toward:

14 Taking care of the needs of veteran teachers
7 Tending to the needs of the district's newer teachers
59 It's about equal
20 Not Sure

77. Which of these best describes what it means to you personally to be a member of a teachers union or association:

31 It provides feelings of pride and solidarity, in addition to the practical benefits
52 It brings practical benefits, not really any more than that
7 It is something that makes you feel uncomfortable
10 Not Sure

78. Other than receiving mailings and notices, how involved and engaged are you in the local union?

6 Very Involved
18 Somewhat Involved
35 Not Too Involved
39 Not At All Involved
2 Not Sure

79. During the time you have been a public school teacher, have you personally filed a grievance against a district or not? [If you have filed more than one, please think about the most recent.] *[Base: Personally Filed a Grievance]*

8 Yes 92 No 1 Not Sure

80. Did it end in your favor, or not? *[Base: Personally Filed a Grievance]*

73 Yes 25 No 3 It is currently in process

81. In general, did you feel the union:

70 Did a good job representing you
27 Could have worked a lot harder
4 Not Sure

82. How much do you know about charter schools?

5 A great deal
18 Quite a bit
39 Only some
32 Very little
7 Nothing at all
* Not Sure

83. As you may know, charter schools operate under a charter or contract that frees them from many of the state regulations imposed on public schools and permits them to operate independently. How much do you favor or oppose the idea of charter schools?

9 Strongly Favor
33 Somewhat Favor
20 Somewhat Oppose
24 Strongly Oppose
14 Not Sure

84. In several districts across the nation, teachers unions are sponsoring and managing charter schools. Do you:

34 Generally support this because it means school policies would be set by the people best qualified to run the school—the teachers and their union
26 Generally oppose this because charter schools are a threat to traditional public schools and the union might make decisions that are not in the best interests of teachers
40 Not Sure

85. How old are you?

2 24 years old or less
8 25 to 29 years
10 30 to 34 years
10 35 to 39 years
23 40 to 49 years
21 50 to 54 years
27 55 or more years

86. Is teaching your first career, or did you work full time in another field beforehand?

71 First career
29 Worked full time in another field beforehand

87. Were either of your parents a public school teacher when you were growing up?

16 Yes 84 No

88. Whether or not they were public school teachers, were either of your parents a member of a union when you were growing up?

35 Yes 65 No

89. Which best describes your school?

14 Inner city
15 Urban
42 Suburban
29 Rural

90. Approximately how many students are in your school?

10 Less than 300
26 300 to 499
36 500 to 999
19 1,000 to 1,999
9 2,000 or more

91. How many of your school's students are African-American or Hispanic?

12 Virtually all
17 Most
46 Some
25 A few or none

92. Approximately what percentage of students at your school are eligible for the free or reduced lunch program? Your best guess is OK.

26 Under 25%
31 25% to 49%
21 50% to 74%
22 75% or more

93. What state do you teach in?
[By region]

17 Northeast
28 Midwest
35 South
20 West

94. As far as you know, which national organization is your district's union or association affiliated with?

1 None
13 American Federation of Teachers (AFT)
71 National Education Association (NEA)
2 Something else
3 AFT and NEA
10 Not Sure

95. Which of the following best describes your own race/ethnicity?

5 African-American or Black
1 Asian or Pacific Islander
4 Hispanic or Latino
88 White or Caucasian
2 Something else

EXHIBIT 7
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

Performance Screens for School Improvement:
The Case of Teacher Tenure Reform in New York City

Susanna Loeb
Stanford University

Luke C. Miller
University of Virginia

James Wyckoff
University of Virginia

May 2014

Thanks to Joanna Cannon, Anne-Marie Hoxie and Keely Alexander at the New York City Department of Education for providing the data employed in this paper and for answering questions about the NYCDOE tenure policy. We appreciate financial support from the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by IES Grant R305A060018. The views expressed in the paper are solely those of the authors and may not reflect those of the funders. Any errors are attributable to the authors.

The effects of public school teacher tenure remain hotly debated, but little understood. Research provides little evidence on the effects of tenure policy choices on teaching quality and, thus, little guidance on how to structure tenure policies. In this paper we examine the effects of a substantial tenure reform in New York City initiated in 2009-10. Below we summarize the key findings from this research, followed by a more detailed discussion of the New York City tenure reforms, our approach to the research, and the findings.

Highlights

- Tenure reforms in NYC led to a substantial drop in the percent of eligible teachers approved for tenure – from 94 percent during academic years 2007-08 and 2008-09, the two years prior to the introduction of the policy, to 89 percent in the first year of the policy (2009-10) and to an average of 56 percent during the three subsequent years.
- The vast majority of eligible teachers who were not approved for tenure had their probationary period extended. The proportion of teachers denied tenure changed only slightly, from two to three percent, following reform.
- Being extended meaningfully increased the likelihood a teacher would transfer across schools or exit teaching in New York City. The probability of transferring was nine percentage points higher and the probability of exiting was four percentage points higher for teachers who were extended compared with teachers in the same school receiving the same principal ratings who were approved for tenure. These differences represent a 50 percent and a 66 percent increase in the probability of transferring and exiting, respectively.
- Extended teachers who transferred or exited were less effective, as measured by principal ratings and value-added, than those likely to replace them. There were 45 percentage points fewer teachers rated as highly effective or effective among all extended leavers than their proxy replacements. In addition, estimated value-added in ELA among extended leavers was 20 percent of a standard deviation lower than among the proxy replacements.
- Schools vary in the proportion of teachers approved, extended and denied tenure. In particular, schools with higher percentages of black students and lower percentages of white students have been more likely to extend and deny teachers for tenure than those with relatively fewer black and more white students. These differences are largely explained by differences in teachers' effectiveness ratings as assigned by principals based on the district-developed Effectiveness Framework. Because extended teachers are more likely to exit, schools with larger enrollments of black students may disproportionately benefit from the reform given that relatively more effective teachers replace extended teachers who voluntarily exit.

Introduction

This paper describes teacher tenure reforms first enacted by the New York City Department of Education (NYCDOE) during the 2009-10 academic year (AY) and the changes in the district's teacher workforce following the reforms. We show that the reforms dramatically changed the proportion of eligible teachers receiving tenure, as well as the career paths of early career teachers, more generally.

Teacher tenure has been controversial since the first tenure provisions were enacted over a century ago. Proponents typically argue that tenure prevents teacher dismissal for political purposes or due to capricious decisions by administrators or politicians. Tenure could guard against dismissal of more experienced, higher paid teachers during periods of tight budgets when school leaders may be more focused on reducing costs while meeting class size requirements than they are on student learning. Tenure does not require schools or districts to retain ineffective teachers but instead provides a due process mechanism to dismiss tenured teachers for cause. Critics, however, argue that the cost of due process does, in practice, lead districts to retain ineffective teachers and as a result tenure not only allows poor teachers to stay in the classroom but also reduces the incentive for teachers to be as effective as they could be. They argue that the due process mechanisms for removing teachers with tenure are so burdensome that they rarely are pursued.

With the availability of large-scale student performance measures linked over time has come clear evidence that teachers vary substantially in their effectiveness at improving student test performance and that these differences can have meaningful effects on students in both the short run and the long run (Chetty, Friedman, & Rockoff, 2012; Rivkin, Hanushek and Kain, 2005; Rockoff, 2004). At least partially as a result, education reforms in the US recently are focusing on improving the quality of teaching through human resource policies such as improved evaluation systems and differentiated pay. Given the controversial nature of teacher tenure, it is not surprising that interest also has increased in changing teacher tenure provisions so that the due process is less onerous and so that school leaders have greater control over their workforce. Yet, the evidence on which to base reform decisions is scarce. We know little about what types of tenure provisions improve the quality of teaching and what types do not. Similarly, we know little about how long the probationary period prior to tenure should be, if there is tenure, in order for school systems to accurately assess teachers' effectiveness so that they can make well informed decisions about tenure.

Part of the reason that we have little evidence on the effects of tenure is that until recently tenure laws have been relatively stable over time and similar, though not the same, across states. New Jersey passed the nation's first teacher tenure law in 1909. Over the next several decades other states adopted similar laws: New York in 1917, California in 1921, and Michigan, Pennsylvania, and Wisconsin in 1937. The state statutes used a variety of synonyms for tenure: continuing contract or service, permanent status, career status, and post-probationary status. Regardless of the terminology, these laws have three main components: tenure requirements, reasons for dismissal, and process for appeals. The first specifies the length of the probationary period after which teachers are eligible for tenure. Employers can dismiss a non-tenured teacher at any time for any reason so long as the decision is neither arbitrary or capricious nor discriminatory, but tenured teachers can only be dismissed for the reasons provided in the law. The third component details the appeals process a dismissed tenure teacher can pursue in an effort to be reinstated. Of the 48 states in which public elementary and secondary teachers are awarded tenure, the minimum probationary period exceeds three years in 11 states (National Council on Teacher Quality, 2012). In most states it is three years, although in a few states, such as California, teachers typically receive tenure with fewer than three years of experience.

The last decade or so have seen substantial change in tenure laws in the US. In 2000, Georgia eliminated due process rights for teachers hired after 1 July 2000, but reinstated these rights three years later. Florida eliminated teacher tenure in 2011. That same year Idaho enacted a law that would have eliminated teacher tenure had it not been repealed by voters the following year. Voters in South Dakota turned back an effort to repeal a 2012 law thereby allowing a law eliminating tenure after 1 July 2016 to take effect. Most recently, North Carolina's governor signed a bill into law that

eliminates teacher tenure by 2018. Though almost all states currently grant tenure, more than half now require meaningful evaluation during the tenure process. As an example, in 2009 only four states used student test performance as a criterion for tenure; by 2012, 20 states did and 25 states require multiple categories for teachers in their evaluation, not just satisfactory and unsatisfactory (National Council on Teacher Quality, 2012). Most recently, the conflicting perspectives on tenure has played out in Vergara v. California, the law suit challenging teacher tenure in California.

A recent reform by the NYCDOE provides an unusual opportunity to learn about the role of tenure in teachers' career outcomes including both strategic retention on the district side and choice-based retention stemming from teachers' decisions. In what follows, we start by describing the reform. We then use data from NYCDOE and the New York State Education Department (NYSED) to provide initial evidence on the magnitude of responses to the reform, concluding with a discussion of the results.

The Teacher Tenure Process in New York City

The criterion for tenure in New York City is that a teacher possesses "significant professional skill and a meaningful, positive impact on student learning." This criterion is not new. However, prior to AY 2009-10 the tenure process in New York City was similar to that in many other large urban districts. The receipt of tenure had become an expectation for nearly all teachers and frequently was based on little evidence of accomplishment. In 2007-08 and 2008-09, well into the period of accountability reforms, 94 percent of all eligible teachers were approved for tenure.

Beginning in 2009-10, New York City changed the tenure review process, infusing more information and increasing the responsibility and accountability of principals to insure that teachers met challenging performance standards. Tenure decisions in 2009-10 were informed by sources of information that had been available previously: classroom observations, evaluations of teacher work products including lesson plans, and the annual rating sheet that principals completed giving teachers a Satisfactory, Doubtful, or Unsatisfactory rating. In addition, tenure decisions in 2009-10 included new student learning measures from the Teacher Data Reports (which included teacher value-added), in-class assessments aligned with the New York State standards, and other evidence of student progress (NYCDOE, 2009).

As in previous years, principals sent recommendations to the superintendent about whether a teacher should be denied tenure, have their probationary period extended or be granted tenure, but starting in 2009-10 principals had to provide a rationale for this decision if the evidence available at the district level suggested either a strong case to approve or deny tenure and this information ran counter to the principal's recommendation. The district provided principals with tenure guidance for teachers for whom there was evidence that performance was particularly strong or weak. For a teacher whose value-added results had been in the lowest 50 percent over the past two years (with a 95 percent confidence interval), who had previously received an Unsatisfactory annual rating, or whose tenure decision had previously been extended, the principal received guidance from the district that the teacher should be considered to have "tenure in doubt". A principal recommendation to extend or approve tenure for these teachers required a supporting rationale for the superintendent to consider in his or her review. The principal received guidance of "tenure likely" for a teacher whose value-added results had been in the highest 50 percent over the past two years (with a 95 percent confidence interval). Principals recommending denying tenure or extending the probationary period for these teachers similarly needed to provide supporting evidence to the superintendent (NYCDOE, 2009).

The process introduced in 2009-10 remained in place in 2010-11 with some notable changes (NYCDOE, 2010). New in 2010-11, principals were asked to evaluate all teachers up for a tenure decision based a four-point effectiveness rating scale (Highly Effective, Effective, Developing and Ineffective) as described in the district-developed Effectiveness Framework.[1] As in the prior year, the evidence for these ratings came from measures of the teacher's impact on student learning such as value-added measures from the Teacher Data Reports, student work products and tests aligned to the New York State standards. Principals also could use evidence from measures of instructional practice coming from their own classroom observations, teacher work products, and the annual rating sheet that principals complete for each teacher.[2] In addition to these sources of information, which were available in the prior year as well, principals in 2010-11 gained information about professional contributions from surveys of students and parents, from measures of attendance, from colleague feedback, and from work products related to the Comprehensive Educational Plan for each school. In contrast to 2009-10, principals in 2010-11 no longer received "tenure likely" or "tenure in doubt" guidance from the district but rather were given flags indicating a "low value add" teacher as an "Area of Concern" and a "high value add" teacher as a "Notable Performance". Low and high value-added scores were defined as in the previous year. Other problematic teacher behaviors flagged as Areas of Concern included: low attendance (defined as exceeding 20 days in the previous two fiscal years), an Unsatisfactory or Doubtful rating on a prior Annual Review Sheet, having been previously extended, having been previously excessed or currently in the Absent Teacher Reserve pool.

The tenure review process for 2011-12 was very similar to that in 2010-11, but with two important changes. As before, teachers were evaluated on impact on student learning, instructional practice and professional contributions. Principals were provided guidance as to the expected (though not required) alignment between the effectiveness ratings they determined using the Effectiveness Framework and their tenure recommendations: Highly Effective and Effective ratings were evidence in favor of granting tenure; a Developing rating, evidence for an extension; and an Ineffective rating, evidence for denying tenure. Additionally, responsibility for producing teacher value-added estimates shifted from the district to the New York State Education Department beginning with 2010-11 and no measures were available for principals to incorporate them into their 2011-12 tenure decisions (NYCDOE, 2011).

The state-provided value-added estimates did inform principals' 2012-13 recommendations. Teachers received a growth score (0-20) that corresponded to a HEDI rating (Highly Effective, Effective, Developing, and Ineffective). No explicit guidance was provided to principals as to how to incorporate these growth ratings into their tenure recommendations. They were only told these ratings are a source of evidence for a teacher's impact on student learning.

Research Questions

Conceptually, the changes in the tenure process could well affect tenure outcomes. As new information on teacher performance becomes available to principals and pressures to be selective in granting tenure increase, the proportion of teachers receiving tenure could decrease. These changes

---

[1] These effectiveness ratings are distinct from the ratings built into the new statewide teacher evaluation system which was not implemented until 3 years later in 2013-14. Although they use the same ratings scale, both the evidence synthesized and the relative weight assigned to the evidence differs between the two.

[2] These sources of evidence were employed in 2009-10 tenure decisions but they were not aggregated in the effectiveness ratings.

could then lead to changes in teachers' choices. Teachers whose likelihood of receiving tenure diminishes may be more likely to leave teaching in the district even if they are not dismissed. Alternatively, some teachers may put more focus on improving the measures of their performance to improve their probability of receiving tenure. Because school principals play a central role in the process and because the teacher workforce differs across schools, we might expect the changes to differ across schools. In keeping with these potential effects, we address the following three research question in this paper:

1. Tenure Decisions – How did tenure rates change following reform?
2. Workforce Composition – Of teachers who become eligible for tenure, how did the composition of those continuing to teach in NYC change following reform?
3. School Differences – How have schools varied in their tenure decisions and the subsequent behaviors of their teachers?

Data

In order to assess the effects of NYCDOE tenure reforms, we must accurately identify teachers eligible for tenure, as well as other teachers potentially affected by the changes. The Tenure Notification System (TNS) tracked the tenure review process for all probationary teachers in New York City public schools between 2007-08 and 2012-13. Each school year, the district made tenure decisions for teachers whose probationary period was scheduled to conclude between November 1st of the current school year and October 31st of the following school year. The probationary period for the 2009-10 cohort, for example, concluded between November 1, 2009 and October 31, 2010. The TNS provided principals with a list of teachers at their school eligible for tenure as well as all official guidance concerning each teacher's job performance prior to the current year (e.g., prior Unsatisfactory annual performance ratings, low attendance, value-added classification, etc.). Principals enter their preliminary and final ratings and recommendations into the TNS and district superintendents make and record final tenure decisions into the system.

We assembled additional information on all teachers, not just those in the TNS, from a variety of sources. NYCDOE provides basic teacher demographic characteristics, the value-added calculations for 2008-09 and 2009-10, the state's value-added calculations for 2011-12 and annual performance ratings used in the tenure review process. We identify teachers' pathways into the teaching profession from state certification records and rosters for the New York City Teaching Fellows program and Teach for America corps members in the New York City region. State certification files provide scores on certification exams. From the College Board we obtain teachers' SAT scores for those teachers who attended a New York public school from 1980 to 2008 or a New York private school from 1980 to 2001. Characteristics of the schools in which teachers teach (e.g., race/ethnicity, free/reduced-price lunch eligibility, AYP status, etc.) come from the annual state-level School Report Cards database and Institutional Master Files and the federal Common Core of Data.

Finally, leveraging data from the NYCDOE Teacher Data Initiative, we observe characteristics of the students taught by specific teachers of grades 4 through 8 mathematics and English language arts (ELA) including demographic and achievement information. We use these data to estimate our own value-added measures of teacher effectiveness using a residuals-based approach controlling for individual student, classroom, and school characteristics. Currently, 2010-11 is the final year for which we can calculate these estimates.

Table 1 provides summary statistics for our analytic sample of teachers receiving a tenure decision in 2010-11 or 2011-12. Just over three quarters of the teachers in our sample are female, approximately 19 percent are black and approximately 17 percent are Hispanic. They have average math and verbal SAT scores of approximately 500 points each. Approximately half of the teachers entered teaching through traditional teacher preparation programs that recommended certification, while 24 percent came through the Teaching Fellows Program, the largest early-entry route serving New York City. These teachers work at schools where 44 percent of students are Hispanic students and 31 percent are black, with 74 percent eligible for subsidized lunch.

Table 1 also includes performance measures for teachers. Recall that principals complete an Annual Rating Sheet for each teacher. Just 2.3 percent of teachers in our sample received an Unsatisfactory rating and one tenth of one percent of teachers received a Doubtful rating, with the remaining 97.6 percent receiving a Satisfactory rating. On the four-point effectiveness rating scale assigned by their principals, most teachers received either a Developing (29 percent) or an Effective (41 percent) rating, while 17 percent received Highly Effective and two percent received Ineffective ratings. Principals provided no effectiveness rating for 11 percent of teachers. Eight percent of teachers had what the district classified as low attendance (more than 20 absences over prior two years), and 12 percent had low value-added.

Results

Tenure Decisions

As described in Figure 1, 94 percent of teachers were approved for tenure during AY 2007-08 and 2008-09, the two years prior to the introduction of the policy. The approval rate dropped to 89 percent in the first year of the policy (2009-10) and averaged 56 percent in the three subsequent years. Virtually all of the decrease in the tenure approval rate resulted in an increase in the percentage of teachers whose tenure decisions were extended, which averaged less than 4 percent prior to the policy, but 41 percent in 2010-11 through 2012-13. The percentage of teachers denied tenure increased marginally following the introduction of the program from an average of two percent pre-policy to three percent post-policy.

Principals have played an important role in the determination of tenure decisions. As shown in Table 2, principal effectiveness ratings using the Effectiveness Framework of teachers are highly predictive of tenure outcomes under the new policy. Ninety-four percent of teachers rated Highly Effective and 83 percent of those rated Effective were approved for tenure. In contrast, less than two percent of those rated Developing and less than one percent of those rated Ineffective were approved. The vast majority (97 percent) of teachers rated Developing were extended, while the vast majority (81 percent) of those rated Ineffective were denied tenure. Given that almost all teachers were approved for tenure prior to the reform, many teachers who would have been approved prior to the reform received a different outcome under the new system.

Tenure decisions also correspond with other teacher performance measures as shown in Table 3. For teachers in tested grades and subjects, value-added estimates track tenure decisions.[3] Teachers denied tenure have math value-added estimates that are a full standard deviation in teacher effectiveness lower than those approved for tenure. On average, extended teachers are 13 percent of a standard deviation in student achievement less effective than the average teacher and 38 percent of a standard deviation less effective than those who are approved. Value-added differences in ELA are smaller but demonstrate the same pattern. Similarly, extended teachers are far more likely to have

[3] We estimate the value-added measures reported in the results section employing a method described in data section.

had prior Unsatisfactory or Doubtful annual performance ratings and to have had Low Attendance than are teachers approved for tenure.

Even though there are substantial differences across the three tenure outcomes in teacher characteristics such as mean value-added estimates and the percent of teachers receiving Unsatisfactory or Doubtful rating or with low attendance, there remains substantial overlap in performance measures among accepted, extended, and denied teachers. For example, as shown in Figure 2, which graphs the distribution of value-added scores for extended and approved teachers, many higher value-added teachers are extended and many lower-value-added teachers approved.

Table 3 also shows patterns between tenure decisions and teachers' background characteristics. While the differences are relatively small, teachers who are approved for tenure have somewhat higher SAT math and verbal scores and teacher certification (LAST) exam scores than those who are extended. Extended teachers, in turn, have somewhat higher test scores than those denied tenure. We find some differences in tenure decisions by pathways as well with New York City Teaching Fellows and teachers entering through Individual Evaluation (IE) less likely to receive tenure than teachers entering the district from college recommending (traditional teacher education) programs.

Overall, the reforms dramatically reduced the percentage of teachers who received tenure, but because most teachers who became eligible for tenure were extended and not dismissed it is unclear a priori whether the reform meaningfully altered the workforce.

Workforce Composition

Changes in the tenure process can affect the quality of teaching by denying tenure to less effective teachers. We found some evidence of this mechanism in Table 3 in that denied teachers had lower value-added in both math and ELA than teachers who were extended or approved. However, even under the new policies, few teachers are dismissed. Larger changes in the workforce instead may come from changes in voluntary turnover, particularly of teachers who are extended or who receive indications that they are likely to be extended.

Extended teachers may voluntarily exit from New York City, creating vacancies which can be filled by more effective teachers. We find some evidence of this phenomenon. As shown in Figure 3, extended teachers were more likely to transfer to other New York City schools and exit from New York City in the year following their decision than teachers who were approved for tenure. Ninety percent of approved teachers return to their schools, while only 75 percent of extended teachers did so.

Being extended meaningfully increases the likelihood of transfers and exits even after controlling for teacher and school characteristics. Table 4 gives the results of regressions with controls for the final principal effectiveness rating of the teachers as well as school fixed effects. The probability of transferring increases by 9 percentage points if the teacher had been extended rather than approved. This represents a *50 percent increase* in the probability of transferring following a tenure decision. Similarly, extended teachers exit NYC at a rate that is 4 percentage points higher than approved teachers, holding other factors constant. This represents a *66 percent increase* in the probability of exiting. These results provide evidence that the new tenure process is having a substantial effect on the composition of the teaching workforce even without substantially increasing the percentage of teachers directly denied tenure.

Among extended teachers, those who remain in the same school have somewhat different measured attributes than those who transfer or exit the system. As shown in Table 5, teachers with

higher academic qualifications, such as teacher certification exam scores, are less likely to stay in the same school than to exit. Extended teachers entering through alternative routes such as the New York City Teaching Fellows program or Teach for America are less likely to remain in the same school than teachers entering through college recommended programs. In contrast, the average value-added estimates of extended teachers who remain in the same school are higher than those who do not, but the sample sizes are smaller for these measures and the differences are not statistically significant at traditional levels.

Are the relatively less effective teachers who are induced to voluntarily leave as a result of tenure reform replaced by more effective teachers? We explore this question by comparing the effectiveness of teachers who were extended and left schools in 2010-11 or 2011-12 with teachers hired at these schools.[4] Teacher effectiveness measures for teachers hired at these schools in 2011-12 and 2012-13 (actual replacements) are unavailable. Rather we employ the effectiveness of teachers hired at these schools in 20008-09 and 2009-10.[5] For each school with an extended leaver, we compare the average effectiveness of extended leavers with that of their proxy replacements, and then average these within school differences across all such schools. In this way we examine the difference in teacher effectiveness between extended leavers and proxy replacements in the typical school.

As shown in Table 6, there are substantial differences in the effectiveness of extended leavers and their proxy replacements. For example, there are 45 percentage points fewer teachers rated as highly effective or effective among all extended leavers than their proxy replacements (14 percentage points Highly Effective and 31 percentage points Effective). Estimated value-added in ELA is 20 percent of a standard deviation higher among the proxy replacements than the extended leavers.[6] Although proxy replacement teachers are estimated to outperform extended leavers in math value-added, this difference is not statistically significant at traditional significance levels, due primarily to relatively few observations (N=158).

From a principal's perspective, these are large effects relative to almost any other intervention they might contemplate. For example, many principals rightly privilege experience when hiring teachers as the value-added of a teacher with six years of experience is estimated to be up to 15 percent of a standard deviation higher than a novice teacher (Atteberry, Loeb and Wyckoff, 2013). Extending the probationary period of teachers with insufficient skills to be approved for tenure and thereby nudging some teachers to leave the school who are then replaced with a new teacher has an effect on teacher effectiveness about the same as the gains of hiring a teacher with six years of experience rather than a novice.

---

[4] Teachers who were hired include both those new to teaching and teachers who transferred from other schools.

[5] The vast majority of teachers with tenure decisions in 2010-11 and 2011-12 began their probationary periods in 2008-09 or 2009-10. We therefore are comparing the extended leavers to other teachers hired under similar circumstances to themselves. We are making the assumption that the teachers hired in 2008-09 and 2009-10 at the schools where an extended teacher left in 2010-11 or 2011-12 have measured effectiveness similar to those teachers who hired at these schools in 2011-12 and 2012-13. We have also created a replacement comparison group of teachers by examining teachers who were hired at these schools from 2006-07 through 2009-10.

[6] Employing the sample of teachers entering schools between 2006-07 and 2009-10 as the proxy replacement comparison group, we estimate the percentage of teachers rated highly effective or effective is 44 percentage points higher for the proxy replacements than the extended leavers. Estimated value-added is 13 percent of a standard deviation higher in ELA and 14 percent of standard deviation higher in math, which are both significant at the 0.06 level.

School Differences

While implementation of the policy may have varied across schools, most schools experienced a substantial change in the percentage of teachers who were approved for tenure under the new policy. More than 70 percent of schools granted tenure to fewer than 80 percent of their teachers following the introduction of the policy as shown in Figure 4. While a cluster of schools approved 100 percent of eligible teachers, most schools approved far less, with another large cluster of schools with between 50 and 70 percent approval.

The variation in approval rates seen in Figure 4 corresponds to some school characteristics, particularly average student attributes, as shown in Table 7. On average, teachers approved for tenure work in schools in which the percentage of white students is nearly twice as large as the schools where teachers were denied tenure. Black students experience the reverse. In schools where teachers are approved for tenure, black students comprise 27 percent of all students, but they comprise 40 percent of students in schools where teachers are denied tenure. The achievement of students in schools where teachers receive tenure is nearly a quarter of a standard deviation better in math and 18 percent of a standard deviation better in ELA than the average achievement in schools where teachers are denied tenure.

Given the strong link between principal effectiveness ratings and tenure decisions shown above, it is not surprising that the pattern of differences in school attributes across principal effectiveness ratings mirror the differences across tenure outcomes as shown in Table 7. For example, the average highly effective teacher works in schools where the percentage of white students is twice as large as it is for the average ineffective teacher. The average ineffective teacher is located in a school with 65 percent more black students than their average highly effective colleague. As is also shown in Table 7, the average ineffective teacher is located in a school where the ELA performance of students is more than a quarter of standard deviation lower and more than 30 percent of a standard deviation lower in math than that of the average highly effective teacher. This suggests that replacing ineffective and developing teachers with a teacher whose performance is closer to the average would disproportionately improve the quality of teaching in schools with higher percentages of black students.

Table 8 describes the relationship between school characteristics and tenure decisions in a multivariate framework controlling for teacher performance measures. When we estimate the model including only the attributes of the students in the school, the percentage of students who are black is the only measure that corresponds to the likelihood of being extended. When teacher attributes are added to the model, they dominate the determination of whether a teacher is extended. The estimate for the percent of black students drops substantially in magnitude such that a 1 standard deviation increase in the percentage of black students (26.4 percentage points) is estimated to increase the likelihood of a teacher being extended by just over 1 percent.

Discussion

Teacher tenure has been a hotly debated issue for decades, but there is surprisingly little research that documents the effects of various tenure policies. This paper examines an unusual change in the tenure policy in New York City as a step toward providing evidence to support the design of teacher workforce policies.

Our analysis documents substantial changes in tenure decisions following the NYC reforms. While almost all eligible teachers received tenure prior to the change, after the reforms a large share of teachers did not receive tenure when they were first eligible, and instead had their probationary

periods extended to provide more opportunity for them to demonstrate the skills necessary for effective teaching and for district decision makers to better assess teachers' performance. Not surprisingly, low-performing and less qualified teachers were more likely to be extended. Teachers in schools with disproportionate shares of black and low-performing students also were more likely to be extended. Our analyses provide some evidence that this differential reflects a uneven distribution of less effective teachers, which is consistent with recent research (Isenberg et al., 2013; Sass et al., 2012), although we cannot rule out differential application of tenure rules. Finally, we found evidence that the tenure policy resulted in additional voluntary attrition of teachers who were extended, as well as additional involuntary dismissal of the small share of teachers who were denied tenure. Among extended teachers, those with lower effectiveness, as measured by principals' ratings, but higher qualifications (e.g. SAT scores) were more likely to leave, potentially further benefiting the teacher workforce. Extended teachers who leave their schools are less effective as measured by principal ratings and value-added estimates than are those likely to replace them. Because teachers with poor effectiveness ratings are more likely to be in schools with higher percentages of black students, these schools are most affected by the policy change and most likely to see attrition of these less effective teachers as a result of the reforms. These schools on average were able to hire more effective teachers to fill these vacancies.

New York City's reforms to the tenure process are still in their early stages. Our results suggest large effects but provide only preliminary evidence because we have not fully ruled out the effects of other factors that may have been at play in the district simultaneously. With additional data a causal analysis will be more feasible and we can address additional questions. While the direct effects of the tenure reforms are felt by teachers facing tenure decisions, the labeling of teachers and increased likelihood of receiving an extension may induce other teachers in the same school, subject, and/or grade to reassess their positions. These processes may encourage principals to reassign teachers across grades and subjects or to reallocate responsibilities in other ways.

Changes in human resource practices including new hiring and evaluation policies have been hallmarks of many recent reforms. While the tenure process has been the subject of continual debate, reforms have been slower and less sustained in this area. In part as a result, research on tenure policies and variety of possible approaches to probationary periods and screening is sparse. Nearly all districts grant some form of tenure based at least in theory on teachers demonstrating proficiency. Yet many districts do only cursory evaluation during the tenure process. As such, adopting tenure reform similar to that presented here may be comparatively easy relative to other much discussed human resource policies that require more controversial policy changes.

References

Atteberry, A., Loeb, S. and Wyckoff, J. (2013). "Do First Impressions Matter? Improvement in Early Career Effectiveness," CALDER Working Paper No. 90, February 2013.

Chetty, Raj, Friedman, John N., & Rockoff, Jonah E. (2012). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Working Paper 17699. Cambridge, MA: National Bureau of Economic Research.

Isenberg, E., Max, J., Gleason, P., Potamites, L., Santillano, R., Hock, H., Hansen, M. (2013). Access to Effective Teaching for Disadvantaged Students. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

National Council on Teacher Quality (2012). *State of the States 2012: Teacher Effectiveness Policies – Area 3: Identifying Effective Teachers*. Washington, D.C.: National Council on Teacher Quality.

New York City Department of Education (2009). *The Tenure Toolkit, 2009-10*. New York: New York City Department of Education.

New York City Department of Education (2010). *The Tenure Toolkit, 2010-11*. New York: New York City Department of Education.

New York City Department of Education (2011). *The Tenure Toolkit, 2011-12*. New York: New York City Department of Education.

Rivkin, Steven G., Eric A., Hanushek, and John F. Kain. (2005). "Teachers, Schools, and Academic Achievement." *Econometrica,* 73(2), pp. 417-458.

Rockoff, Jonah E. (2004). The impact of individual teachers on student achievement: evidence from panel data. *American Economic Review Papers and Proceedings, 94*(2), 247–252.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics*, *72*(2), 104-122.

Figures

Figure 1. Percentage of Teacher Tenure Cases by Tenure Outcome 2007-08 to 2012-13

Figure 2. Distributions of Teacher Value-Added of Approved and Extended Teachers, Math and ELA, 2010-11

Figure 3. Location of Teachers in Year Following Tenure Decision, by Tenure Outcome, 2010-11 and 2011-12

Figure 4. Distribution of School Proportion of Tenure Cases Approved 2009-10 through 2012-13

Notes: Includes only schools with at least four tenure decisions over the period (81 percent of all schools).

Tables

Table 1. Descriptive Statistics for the Analytic Sample, 2010-11 and 2011-12

Variable	Obs.	Mean	Std. Dev.
Tenure Outcome (%)			
Approve	9,161	56.97	49.51
Extend	9,161	40.04	49.00
Deny	9,161	2.99	17.03
Teacher Attributes (% unless otherwise noted)			
Female	9,129	75.53	
Black	8,139	18.64	
Hispanic	8,139	16.89	
SAT math	4,236	499.01	103.00
SAT verbal	4,236	502.00	99.43
Preparation Path (%)			
College recommended	9,084	49.98	
Teaching Fellow	9,084	23.83	
TFA	9,084	3.64	
Individual evaluation	9,084	7.63	
Temporary license	9,084	4.60	
Student Attributes (aggregated to school)			
Hispanic (%)	8,961	44.46	25.32
Black (%)	8,961	30.81	26.42
Free lunch (%)	7,894	74.43	22.20
Reduced lunch (%)	7,894	4.25	4.15
Mean ELA score (z-score)	6,530	2.89	44.16
Mean Math score (z-score)	6,530	1.46	46.85
Teacher Performance Measures (%)			
U rated	9,161	2.33	
D rated	9,161	0.14	
Principal Final Effectiveness Ratings			
Ineffective	9,161	2.22	14.72
Developing	9,161	28.85	45.31
Effective	9,161	41.10	49.20
Highly Effective	9,161	16.97	37.54
No Rating	9,161	10.86	31.12
Low attendance	9,161	7.53	26.39
VAM ELA	1,052	-0.06	1.03
VAM Math	670	-0.06	1.09
NYC VAM low	1,101	11.99	
NYC VAM high	1,101	8.08	

Table 2. Tenure Decision Outcome by Principal Final Effectiveness Rating, 2010-11 to 2012-13

	Ineffective (%)	**Developing (%)**	**Effective (%)**	**Highly Effective (%)**	**None (%)**
Approve	0.7	1.8	82.7	93.9	53.7
Extend	18.2	96.6	17.1	6.1	41.4
Deny	81.1	1.6	0.2	0.0	4.9
N	302	3,820	5,568	2,006	1,384
% teachers	2.3	29.2	42.6	15.3	10.6

Table 3. Attributes of Teachers by Tenure Outcomes, 2010-11 through 2012-13[a]

Tenure	Value Added		U Rated	D Rated	Low Attd	SAT		LAST	Preparation Route (%)[b]			
Decision	*ELA*	*Math*	**(%)**	**(%)**	**(%)**	*Math*	*Verb*	**Exam**	*Coll Rec*	*NYCTF*	*TFA*	*Ind Eval*
Approve	0.081	0.248	5.7	22.2	37.1	505	505	257	59.9	49.5	60.2	55.0
Extend	-0.138	-0.129	52.1	66.7	56.2	490	494	254	37.8	47.2	38.9	40.7
Deny	-0.115	-0.740	42.2	11.1	6.7	469	490	248	2.4	3.2	0.1	4.3
Total	-0.009	0.070	100.0	100.0	100.0	498	500	255	100.0	100.0	100.0	100.0

[a] Means of teachers approved exceed those of teachers extended at a p-value of 0.05 or lower for all attributes. The means of teachers extended exceed those of teachers denied at a p-value of 0.05 or lower for all variables except ELA value-added and verbal SAT.
[b] The tenure approval rate is lower for teachers prepared through the NYCTF and IE preparation routes than those from CR programs at p-values of .01 or lower. There is no statistical difference between CR and TFA.

Table 4. Determinants of Teacher Disposition in Year Following Tenure Decision, 2010-11 and 2011-12

Variables	(1) Transfer	(2) Transfer	(3) Transfer	(4) Exit	(5) Exit	(6) Exit
Extend	0.145**	0.124**	0.087**	0.057**	0.055**	0.040**
	(15.21)	(13.04)	(6.06)	(9.38)	(9.07)	(4.32)
Student Attributes						
Mean Math score	-0.024			0.016		
	(-0.68)			(0.68)		
Mean ELA score	-0.024			-0.019		
	(-0.64)			(-0.82)		
Black (%)	0.113*			0.042*		
	(4.24)			(2.46)		
Hispanic (%)	0.066~			0.075**		
	(2.35)			(4.21)		
Free lunch (%)	-0.099**			-0.085**		
	(-3.12)			(-4.23)		
Reduced lunch (%)	-0.307*			-0.187*		
	(-2.33)			(-2.23)		
Principal Final Effectiveness Rating						
Ineffective			0.285*			0.110*
			(4.24)			(2.54)
Developing			0.071**			0.026*
			(3.58)			(2.02)
Effective			0.030*			0.007
			(2.13)			(0.74)
Missing			0.045*			0.013
			(2.45)			(1.11)
Constant	0.142**	0.135**	0.111**	0.064**	0.037**	0.031**
	(4.88)	(24.56)	(9.60)	(3.48)	(10.52)	(4.22)
School Fixed Effect		X	X		X	X
Observations	6,351	8,855	8,855	6,351	8,855	8,855

Notes: T-statistics in parentheses. ** p<0.01, * p<0.05, ~p<0.1

Table 5. Attributes of 2011 and 2012 Extended Teachers by Disposition in the Following Year

Attrition	Value Added		U Rated	D Rated	Low Attd	SAT		LAST	Preparation Route (%)			
Status	*ELA*	*Math*	**(%)**	**(%)**	**(%)**	*Math*	*Verb*	**Exam**	*Coll Rec*	*NYCTF*	*TFA*	*Ind Eval*
Same School	-0.091~	-0.090	4.0~	0.2**	10.7	491	495	253**	77.5	70.9**	53.3**	78.8*
Transfer	-0.355	-0.421	2.7	0.2	11.2	482	486	253	16.3	15.6	9.0	17.7
Exit	-0.332	-0.145	2.9	0.0	9.1	530	539	267	6.2	13.6	37.7	3.5

Notes: ** p<0.01, * p<0.05, ~ p<0.10. For Value-Added, U Rated, D Rated, Low Attendance, SAT and LAST Exam, significance levels denote significant differences between the values of these variables for Extended teachers who remain in same school and those who either transfer or exit. For Preparation Routes, significance levels denote differences between designated route and College Recommended.

Table 6. Mean School Difference in Teacher Effectiveness Measures between Proxy Replacement and Extended Leavers in Schools with Extended Leavers, 2010-11 and 2011-12[a]

	Principal Final Effectiveness Rating (%)				Value-Added	
Extended Leaver Status	*Highly Effective*	*Effective*	*Developing*	*Ineffective*	*ELA*	*Math*
All Extended leavers	14.34***	30.7***	-36.45***	1.37*	0.197**	0.119
Extended transfers	11.97***	30.16***	-34.53***	1.14	0.127	0.181*
Extended exiters	16.15***	27.55***	-33.24***	1.72	0.298*	0.037

Notes: [a] Proxy replacement teachers are all teachers hired at the school in 2009 and 2010. Only schools with an extended leaving teacher in 2011 or 2012 included in all comparisons. Positive values indicate on average within schools average value for replacement pool exceeds that for the Extended leavers. Comparing extended leavers to proxy replacements *** p<0.001, ** p<0.01, * p<0.05.

Table 7. Attributes of the Students in Teacher's School by Tenure Decision and Principal Effectiveness Rating, 2010-11 and 2011-12

	White (%)	Hispanic (%)	Black (%)	Home Lang Eng (%)	Free Lunch (%)	Reduced Lunch (%)	Math Achieve (z-score)	ELA Achieve (z-score)
Tenure Decision [a]								
Approve	13.8	44.4	27.4	56.6	72.3	4.4	0.081	0.086
Extend	8.9	44.6	35.1	60.3	77.3	4.1	-0.066	-0.042
Deny	7.1	43.5	39.6	63.3	77.8	4.2	-0.152	-0.093
Principal Effectiveness Rating [b]								
Highly Effective	16.4	42.8	24.1	56.5	69.2	4.8	0.184	0.181
Effective	12.1	45.3	29.9	57.2	74.6	4.2	0.007	0.019
Developing	8.4	45.0	35.3	60.8	78.1	4.1	-0.068	-0.046
Ineffective	7.2	42.4	39.9	62.7	77.7	4.6	-0.161	-0.102
No rating	12.3	42.7	31.0	57.4	71.3	4.1	0.055	0.073
Total	11.7	44.5	30.8	58.3	74.4	4.2	0.015	0.029

Notes: [a] Extended teachers work in schools with different student attributes than approved teachers (p-value less than 0.01 for all attributes except the percentage of Hispanic students). Teachers denied tenure work in schools with different attributes than teachers who are extended with respect to the percentage of students who are black, the percentage whose home language is not English and mean student math scores (p-value less than 0.05). Differences in other student attributes are not significantly different from zero.

[b] Teachers rated ineffective work in schools with different student attributes than teachers rated effective or highly effective (p-value less than 0.01 for all attributes except the percentage of Hispanic students and the percentage eligible for reduced-price lunch). Teachers rated developing work in schools with different student attributes than teachers rated effective or highly effective (p-value less than 0.01 for all attributes except the percentage of Hispanic students).

Table 8. Determinants of Whether Teacher is Extended Relative to being Approved, 2010-11 and 2011-12

	(1) **Extended** **(=1)**	(2) **Extended** **(=1)**
Student Attributes		
Mean Math score	-0.096	-0.073~
	(-1.41)	(-1.77)
Mean ELA score	-0.010	0.021
	(-0.14)	(0.49)
Black (%)	0.211**	0.048~
	(-4.41)	(1.80)
Hispanic (%)	0.032	-0.008
	(-0.62)	(-0.27)
Free lunch (%)	0.012	-0.041
	(-0.20)	(-1.13)
Reduced lunch (%)	-0.066	-0.043
	(-0.26)	(-0.28)
Teacher Attributes		
Low Attendance		0.066**
		(3.84)
Unsatisfactory Rated		0.101**
		(2.85)
Doubtful Rated		-0.125
		(-0.75)
Principal Final Rating		
Ineffective		0.867**
		(25.61)
Developing		0.906**
		(95.62)
Effective		0.100**
		(8.72)
No rating		0.334**
		(12.95)
Constant	0.340**	0.081*
	(-6.12)	(2.45)
Observations	6,351	6,351
R-squared	0.033	0.613

Notes: T-statistics in parentheses. ** p<0.01, * p<0.05, ~ p<0.1

EXHIBIT 8
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

Why are most teachers rated effective when most students test below standards?

On Board Online • December 16, 2013

By Cathy Woodruff
Senior writer

Here's a word problem that could stump even the savviest student of Common Core-aligned mathematics:

Less than one-third of New York students passed the state math and English Language Arts tests they took in April. Yet, more than 90 percent of the state's teachers were rated effective or highly effective under the state's new Annual Professional Performance Review (APPR) rating system. Explain.

It's a head-scratcher, all right. How can New York's teachers possibly be so effective if their students are struggling so mightily to meet the state's new academic standards?

Even a more sophisticated analysis – limiting the sample of teachers to those in the elementary and middle school classrooms where students took state exams in April, and limiting the ratings solely to the 20 points tied to those state test results – still reveals a sharp apparent contradiction.

More than 83 percent of the teachers in grades 4-8 were rated effective or highly effective on the portion of their Annual Professional Performance Reviews (APPR) tied to their students' test scores. But just 31 percent of students who took ELA and math tests met the new standards for proficiency on each of them.

How is that possible? Isn't a high level of teacher effectiveness supposed to correlate with high student achievement? Isn't this supposed to be an accountability system?

According to state Education Commissioner John B. King Jr., trying to resolve the apparent paradox of good teacher ratings despite disappointing test scores for their students is a lot like the folly of trying to compare apples to oranges. Students are being tested on their mastery of the standards, but teachers are not actually being evaluated on their students' level of mastery of the standards. Rather, they are being evaluated, in part, on their students' growth in mastering the standards.

That's why teacher scores can be high while student scores are low, educators say.

"I don't think the two are connected at all – at least as the system currently is set up," said Herricks School District Superintendent Jack Bierwirth, who serves on the Metrics Work Group of the State Education Department's APPR Task Force.

But if some New Yorkers were puzzled by the sunny teacher evaluation scores, it would be hard to blame them. After all, they were told repeatedly by state leaders that a teacher evaluation system would be instrumental in improving academic achievement and holding teachers accountable for student learning. They also were told that students' performance on state tests would be a strong, objective indicator of their teachers' effectiveness.

"The new statewide evaluation law sets clear standards for measuring educators based on how our students are performing in the classroom," Gov. Andrew Cuomo declared when he announced a March 2012 agreement with legislative leaders "to put the governor's new groundbreaking teacher and principal evaluation system into law."

Despite that rhetoric, New York's new APPR system does not draw anything close to a straight line between student achievement and teacher and principal evaluation ratings.

First, it must be noted that teachers' overall or "composite" APPR ratings give far more weight to other factors, such as classroom observations and local measures of student learning, than they give to the portion linked to state test results.

And while it's often said that 20 percent of a teacher evaluation is based on state test scores, it would be more accurate to say that portion is based on the degree of change in student test scores. That component is derived from a calculation designed to determine how much a student has improved as a result of a teacher's instruction that year. Performing the growth calculation requires comparisons with prior test performance.

Estimating the student growth component was especially tricky this year because this year's tests measured students against the new Common Core standards, while state tests in previous years were designed to measure performance based on standards set in 2005. That's why the State Education Department sent out a flurry of charts, Excel worksheets, tables and guidance documents in August. The tools were intended to help administrators place old and new student test scores on a common scale so administrators could compare them.

Without comparisons, raw test results are virtually worthless for judging teacher performance, said Bierwirth, the Herricks superintendent. A 2013 score, alone, "doesn't take into account where students started. It only describes where they ended up," he noted.

"I do think the effort to measure a teacher's value, based on what they contribute to a student's learning, is the right direction," Bierwirth added, but he is critical of the metric gymnastics now being used to calculate student growth for use in APPR formulas. "I believe the teacher evaluation system is, as it is now set up, highly flawed and not a terribly good measure of effectiveness," he said.

The complexity and the lack of clear connection between the test scores and APPR ratings is what can make it so hard for policy makers, including school board members, to explain how, exactly, the new system makes schools more accountable for results.

Aaron M. Pallas, a professor of sociology and education with Teachers College at Columbia University, has doubts about how precise educators can expect APPR to be in diagnosing an individual teacher's impact on academic achievement. He says there are just too many other variables in play, including groundwork laid by teachers in earlier grades and whatever is going on in a student's home life.

"It's really hard to isolate the contribution of one teacher to a cumulative level of performance," Pallas said. "I think one recommendation would be to forego some of the false sense of quantified process that APPR has created. All of the components are things that I think are a bit fuzzy. Yet, we are adding them up and treating them as though the result is not fuzzy."

Pallas and other observers say it's also likely that this year's strong overall teacher effectiveness ratings were bolstered by positive ratings for classroom observations and other locally-developed criteria, which were crafted amid concerns about the unpredictable impact of state test scores.

Again, political rhetoric played a role in forming perceptions about the aims and hazards of APPR. For instance, Gov. Cuomo issued a March 2011 news release that touted a statewide teacher evaluation system as an alternative to the "so-called 'last in, first out' seniority policy," which he said "lacks objectivity by maintaining teachers simply based on years of service without factoring classroom effectiveness, performance or need."

"I think that the way the state framed it put too much emphasis on the APPR process as a way to identify ineffective teachers who ought to be drummed out," said Pallas. "In some cases, districts thought they already knew who the good and bad teachers were."

Attorney Howard Goldsmith of the Harris Beach law firm has coined a phrase to describe the problems with perceptions about APPR and school reform in general. He calls it "the disconnect gap."

Writing on the Harris Beach municipal affairs blog, nymuniblog, Goldsmith said that low student scores and high teacher scores are emblematic of a broader problem in which the elements of New York's education reform operations don't work together in a way that's clear.

Better communication could help, Goldsmith said, but he argues that a solution that restores faith in educational reform will require more substantial action.

His suggestions include extending and coordinating the multiple timelines for implementing Common Core standards, new curriculum and tests and APPR, with common benchmarks and transition dates for the various initiatives. He also suggests simplifying the APPR process and removing entirely the second component, which relies on locally determined measures of student achievement.

"Closing the disconnect gap will require adjustments in the actual implementation of the reform agenda initiatives, not just some positive but minimal changes in state testing policies," Goldsmith wrote in nymuniblog. "To close the disconnect gap, adjustments must be taken to properly align the implementation plans of the respective reform agenda elements, making them connected in a logical and easy-to-understand common-sense fashion."

Send this page to a friend

Show Other Stories

EXHIBIT 9
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

THE BOTTOM LINE

Few teachers across New York State earned low ratings last year

by Geoff Decker on October 22, 2013

Teachers who were worried that the state's new evaluation rules could put them at risk of being fired can exhale now. Almost no one was rated ineffective in the first round of ratings under the new rules, state education officials announced today.

Preliminary Statewide Composite HEDI Results: Teachers

HEDI Summary (Teachers)
126,849 (3,920 no composite reported – 3.1%)

HEDI Rating	STATE-WIDE	
Highly Effective	49.7%	91.5%
Effective	41.8%	
Developing	4.4%	
Ineffective	1.0%	
Total	96.9%	

*Note: This summary reflects the data that were reported to the Department by Districts and BOCES as of the 10/18/2013 deadline.

EngageNY.org 5

Just 1 percent of teachers across the state — excluding New York City — were rated ineffective last year, according to the data, Another 4 percent were rated "developing," which signals that teachers should receive additional support.

Fully half of teachers earned the state's highest rating, "highly effective," and another 42 percent were deemed "effective."

The new evaluation system, unveiled in conjunction with new standards for students, was meant to distinguish teacher quality and resolve the disconnect between teachers' almost uniformly high ratings and the state's low college readiness rate.

That did not happen this year. While 92 percent of teachers were highly effective or effective, just 31 percent of students in the state were deemed to be on track for college and careers.

State education officials said the disconnect now shows that tougher academic standards do not prevent teachers from demonstrating excellence, despite what some teachers had feared. "The results are striking," Commissioner John King said in a statement. "The more accurate student proficiency rates on the new Common Core assessments did not negatively affect teacher ratings."

Preliminary rating data the state released over the summer suggested that more teachers were in danger of earning low ratings. That data reflected only the 20 percent of ratings that are based on state "growth scores" and only the fifth of teachers who work in tested grades and subjects. It showed that 6 percent of the teachers were ineffective and another 6 percent were highly effective.

Districts can move to fire teachers who earn "ineffective" ratings two years in a row under the state's new evaluation rules, written into law in 2010 as part of the state's efforts to win funding in the federal Race to the Top competition.

New York City teachers are not included in the new data because the city did not have a teacher evaluation system in place last year because the city and teachers union were unable to agree on a plan, despite pressure from Gov. Andrew Cuomo. Some city teachers did receive growth scores, which showed them outperforming teachers in the rest of the state.

In the rest of the state's roughly 700 school districts, teachers were evaluated according to multiple measures that were split up among standardized state test scores, tests chosen locally, and observations from principals and other administrators. Some districts, such as Syracuse, also factored student surveys into teachers' scores.

Districts had to compile the different components into single composite scores for each teacher and submit them to the state by last week.

The state's presentation, made at this morning's Board of Regents meeting, did not break down the composite ratings by the various measures that make up the scores. King said analyzing the evaluation subcomponents to understand why the state's growth measure did not reflect the ultimate results would be a next step for his department.

Neither did the state's presentation break the ratings down by district. Early indications suggest that scores could range widely across districts: In Syracuse, which released its scores earlier this month, just 60 percent of teachers received the higher ratings, and the teachers union is planning widespread appeals.

© 2014 Chalkbeat

EXHIBIT 10
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

Cornell University
ILR School

Cornell University ILR School
DigitalCommons@ILR

Scheinman Institute on Conflict Resolution

Scheinman Institute on Conflict Resolution

5-2014

APPR Teacher Appeals Process Report

Alexander Colvin
Cornell University, ajc22@cornell.edu

Sally Klingel
Cornell University, slk12@cornell.edu

Simon Boehme
Cornell University, sjb334@cornell.edu

Follow this and additional works at: http://digitalcommons.ilr.cornell.edu/icrpubs

Part of the Educational Assessment, Evaluation, and Research Commons

This Article is brought to you for free and open access by the Scheinman Institute on Conflict Resolution at DigitalCommons@ILR. It has been accepted for inclusion in Scheinman Institute on Conflict Resolution by an authorized administrator of DigitalCommons@ILR. For more information, please contact jdd10@cornell.edu.

APPR Teacher Appeals Process Report

Abstract
[Excerpt] The Annual Professional Performance Review (APPR) is the new teacher evaluation system adopted by New York State in 2012. Through APPR, each New York State teacher's performance is evaluated annually. If a teacher is rated Ineffective, he or she must take part in a Teacher Improvement Plan (TIP). If a teacher is rated Ineffective for two consecutive years, the teacher may be dismissed even if that teacher has tenure. Given these potential consequences, the ability to appeal APPR ratings and how those ratings are conducted has been a major issue for teachers and their unions. Under New York Education Law 3012-c, which establishes APPR, each school district negotiates its own APPR procedure with its local teachers union, including any procedures for appealing the performance review. This report examines the APPR appeals procedures established by school districts in order to investigate the following: Which aspects of the APPR process can teachers appeal? Who has the fi nal say in that appeals process? How much time do appeals processes take? Can teachers appeal APPR issues through the regular contractual grievance-arbitration procedure? This report addresses these questions by analyzing APPR appeal procedures for all New York State school districts. The data analyzed was gathered by coding the provisions of the APPR appeal procedures, which are publicly available on the New York State Department of Education website (1).

Keywords
Annual Professional Performance Review, APPR, teacher evaluation, New York State

Disciplines
Educational Assessment, Evaluation, and Research

Comments
Suggested Citation
Colvin, A., Klingel, S., & Boehme, S. (2014). *APPR teacher appeals process report* [Electronic version]. Ithaca, NY: Cornell University, ILR School, Scheinman Institute on Conflict Resolution.

Required Publisher Statement
© Cornell University. Reprinted with permission. All rights reserved.

This article is available at DigitalCommons@ILR: http://digitalcommons.ilr.cornell.edu/icrpubs/7

ILR School
ADVANCING THE WORLD OF WORK
Bargaining For Better Schools
Scheinman Institute on Conflict Resolution

APPR Teacher Appeals Process Report

BY: ALEX COLVIN, SALLY KLINGEL, AND SIMON BOEHME

MAY 2014

The Annual Professional Performance Review (APPR) is the new teacher evaluation system adopted by New York State in 2012. Through APPR, each New York State teacher's performance is evaluated annually. If a teacher is rated Ineffective, he or she must take part in a Teacher Improvement Plan (TIP). If a teacher is rated Ineffective for two consecutive years, the teacher may be dismissed even if that teacher has tenure. Given these potential consequences, the ability to appeal APPR ratings and how those ratings are conducted has been a major issue for teachers and their unions. Under New York Education Law 3012-c, which establishes APPR, each school district negotiates its own APPR procedure with its local teachers union, including any procedures for appealing the performance review. This report examines the APPR appeals procedures established by school districts in order to investigate the following: Which aspects of the APPR process can teachers appeal? Who has the final say in that appeals process? How much time do appeals processes take? Can teachers appeal APPR issues through the regular contractual grievance-arbitration procedure? This report addresses these questions by analyzing APPR appeal procedures for all New York State school districts. The data analyzed was gathered by coding the provisions of the APPR appeal procedures, which are publicly available on the New York State Department of Education website (1).

Under APPR, there are four possible ratings: Highly Effective, Effective, Developing, and Ineffective. Each rating system is based out of 100 points, 40 of which must be based on measures of student achievement. The remaining 60 points may come from classroom observations or other locally determined evaluation methods. In the student achievement section, 20 points must be composed of state-developed measures of student growth and the other 20 points are based on locally selected measures of student achievement. Based on the accumulated points, each teacher will receive his or her rating of Highly Effective, Effective, Developing, or Ineffective.

The APPR procedure plays an important role in employment decisions, particularly because two consecutive Ineffective ratings immediately subject New York State teachers to an expedited 3020-a hearing for potential dismissal. The 3020-a process provides for an expedited hearing before a mutually-selected hearing officer. Within 60 days of the initial meeting, the hearing officer must provide a final decision to the commissioner, the employee, and the employing board who then implements the decision. In New York City however, teachers with two consecutive Ineffective ratings are first assigned an independent observer to assess the teacher in the classroom before the teacher is subject to the 3020-a process.

What are teachers not able to appeal

Under the new APPR system, each New York State school district locally negotiated its own APPR agreement, establishing the review process and providing for an appeals procedure.

***Table 1:* What Are Teachers Not Able to Appeal?**

	Tenured Teachers	Non-tenured Teachers
Improvement Plan	1.3%	29.8%
Implementation of Improvement Plan	1.1%	35%
Appeal depending on rating:		
Ineffective Rating	0%	26.5%
Developing Rating	14.3%	46.7%
Effective Rating	96.4%	96.6%

N=688

From district to district, there are differences in what tenured and non-tenured teachers are able to appeal and what they are explicitly prohibited from appealing as described in Table 1. Although many contracts set out these rights explicitly, others do not explicitly state what aspects of the APPR may be appealed. In general, contracts are more likely to limit the types of appeals that non-tenured teachers may make, whereas tenured teachers can more often appeal all aspects of the APPR process. Few procedures restrict tenured or non-tenured teachers from appealing the substantive or the procedural aspects of the review itself. The majority of procedures also allow tenured and non-tenured teachers to appeal the provisions of an improvement plan. Appeals concerning the implementation of improvement plans are also allowed for the majority of tenured teachers. However, just less than half of procedures allow appeals concerning the implementation of an improvement plan by non-tenured teachers.

Some APPR procedures limit the teacher's ability to appeal a rating depending on the level of rating being challenged. In addition to always allowing ratings of Ineffective to be appealed, most procedures for tenured teachers also allow a rating of Developing to be appealed. However, 46.7% of agreements bar non-tenured teachers from appealing a Developing rating. Most procedures prohibit a rating of Effective to be appealed by either tenured or non-tenured teachers.

APPR appeal determination

Appeals procedures vary in complexity, including between one and five steps. Although each step involves a different decision maker, the decision maker in the final step is particularly important in ultimately deciding the outcome of the appeal. Table 2, displays the range of final decision makers. Superintendents are the most common final decision maker, serving this role in 77% of the APPR agreements. The next most common type of final decision maker is a panel, jointly appointed by the district administration and the teacher or union. Overall, 15% of procedures feature this type of panel as the final decision maker for APPR appeals. Only 2% of procedures have an arbitrator who makes the final decision. In terms of who serves as the final decision maker in the appeals process, there is little difference in procedures for tenured and non-tenured teachers.

Table 2: **Final Decision Maker for APPR Appeals Process**

	Tenured Teachers	Non-tenured Teachers
Superintendent	77%	76%
Panel	15%	14%
Arbitrator	2%	2%
School Board	1%	1%
External Evaluator	1%	1%
Original APPR Rater	<1%	2%
	N=688	N=445

APPR agreements set time limits for the various steps in the appeals process. Because each step is typically assigned a specific time limit, combining the limit for each step in the procedure gives an overall length of time to complete the full appeals process. Although the overall time limit for completing all steps of the appeals process is an average of 64 days for tenured teachers and an average for 63 days for non-tenured teachers, there is a wide range across districts for these maximum time limits, as shown in Table 3 Half of the procedures for tenured teachers have overall time limits between 44 and 80 days, and 80% of all procedures have overall time limits between 29 and 98 days.

Table 3: **Maximum Time Limits for APPR Appeals Process (in days)**

	Mean	Range (10th to 90th percentile)
Tenured	64	29-98
Non-tenured	63	30-101

Exclusivity clause

How do the APPR appeals processes interact with the grievance-arbitration procedures established in collective bargaining agreements? For violations of the collective bargaining agreement, all New York State teacher contracts have a grievance procedure. Most districts chose to create an APPR appeals process that does not allow the use of the collective bargaining agreement grievance procedure. As shown in Table 4, 66% of APPR procedures include a type of exclusivity clause that prevents teachers from using the collective bargaining agreement grievance procedure to appeal APPR ratings.

By contrast, 14% of APPR appeals procedures explicitly include language that allows teachers to appeal their rating through the grievance procedure after exhausting the APPR appeals procedure. Another 5% of APPR appeals procedures allow procedural, but not substantive, issues from the APPR to be appealed through the collective bargaining agreement grievance procedure after the APPR appeals procedure is completed. In only one district, Auburn City School District, was the collective bargaining agreement grievance procedure itself used as the appeals process to resolve APPR appeals. The remaining 15% of procedures fail to explicitly state whether or not a teacher may appeal APPR issues through the collective bargaining agreement grievance procedure.

Table 4: Relationship of APPR to Collective Bargaining Agreement (CBA) Grievance Procedure	
Procedural or Substantive APPR Appeals Cannot Use CBA Grievance Procedure (Exclusivity Clause)	66%
Procedural or Substantive APPR Appeals Can Use CBA Grievance Procedure (Exclusivity Clause)	14%
Only Procedural APPR Appeals Can Use CBA Grievance Procedure	5%
No Statement Whether APPR Can Use CBA Grievance Procedure	15%

N=688

Second consecutive ineffective rating

Following a second consecutive Ineffective rating, a tenured teacher may be fired after a 3020-a hearing. Meanwhile, non-tenured teachers can be fired after the second Ineffective rating without going through a 3020-a hearing. Given these potential consequences, some districts have established additional appeals procedures for a second consecutive Ineffective rating. Our research finds that nearly 10% of contracts have this type of special appeals process if the teacher has received a second rating of Ineffective. Most of these special appeals processes are specifically for tenured teachers (6.8% of all contracts), but some of these processes are open to non-tenured teachers (2.6% of all contracts). These special appeals processes occur before and are separate from the 3020-a hearing.

Table 5: Final Decision Maker in Second Consecutive APPR Appeals Process	
Arbitrator	64%
Superintendent	21%
Panel	9%
School Board	3%

N=67

As shown in Table 5, a majority of these special appeals processes use arbitration as the final step in decision making. The second most common type of final decision maker is the superintendent, followed by a joint panel. The average maximum time limit for these special appeals processes is 55 days.

Conclusion

The APPR evaluation system has major implications for school districts and teachers. The possibility of being terminated following a second consecutive Ineffective rating, even if tenured, makes evaluations more important to teachers. As a result, procedures for appealing APPR ratings are vital in providing due process. This report provides an overview of the major characteristics of APPR appeals procedures that school districts have established. In most school districts, teachers are able to appeal the process and substance of APPR ratings, the contents of an improvement plan, and to challenge both Ineffective and Developing ratings. Generally, tenured teachers have broader rights to appeal their APPR evaluation than non-tenured teachers. However, it should be noted that prior to the APPR procedure, non-tenured teachers generally could not appeal school administration evaluations of their performance. So, the APPR appeals procedures are often an upgrade of due process protections for non-tenured teachers.

When an APPR appeals procedure is established, the final decision maker is most often the superintendent, unlike in collective bargaining grievance procedures where the final decision maker is almost always an arbitrator. Some APPR appeals procedures use alternative final decision makers, particularly panels mutually agreed upon by the school district administration and the teacher or union. These panels may be related to existing teacher peer appraisal and development process and will be studied further in future reports.

Cornell University
ILR School

Bargaining for Better Schools (BBS) is an initiative of the ILR School at Cornell University through the Scheinman Institute on Conflict Resolution and the Worker Institute. These publications are free for public reproduction with proper accreditation. For more information on BBS, our past publications, and future research, please visit: www.ilr.cornell.edu/bbs.

Note: "N" stands for the number of contracts
(1) School district APPR appeals procedures can be found at: http://usny.nysed.gov/rttt/teachers-leaders/plans/.

A special thank you to research assistance provided by Honore Johnson, Abigail Frey, Alexandra Reinhardt, Micaela Lipman, and Molly Beckhardt.

EXHIBIT 11
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

TE² Community Brief ▪ May 2012

Earned, Not Given:
Transforming Teacher Tenure

COMMUNITIES for TEACHING EXCELLENCE

Tenure should be considered a significant milestone for teachers who have consistently demonstrated effectiveness and commitment. Unfortunately, in most states tenure is often awarded automatically after a teacher has been in the classroom for two or three years.
National Council on Teacher Quality, 2011

Tenure for public school teachers lies at the center of the current debate on education reform. Opponents believe tenure represents all that is wrong with the current education system, protecting ineffective and unprofessional teachers by giving them lifetime employment without regard for how well they perform. Supporters of the current system argue that teachers must be protected from arbitrary dismissal and undue political interference in their work.

The truth lies somewhere in the middle: tenure need not be abolished, but it must be transformed.

The majority of states and school districts grant tenure to teachers after only two or three years in the classroom and without regard for how well teachers actually perform or how much their students learn. Not surprisingly, there is growing concern that current tenure policies are shortchanging students and jeopardizing opportunities to close our nation's achievement gap. With numerous studies demonstrating that teaching quality has the greatest impact on student achievement—more than any other factor in the school—there are increasing calls to reform current policies so that only the most effective teachers receive tenure and remain in the classroom (Rivkin, et al.).

This brief examines the weaknesses of current tenure systems and discusses promising strategies for addressing these shortcomings. It also highlights some of the reforms that states and school districts are instituting to increase the quality of teachers and to ensure that all students have access to effective teaching. Tenure practices must be transformed so that they serve the best interests of students, while still supporting teachers.

A SYSTEM IN NEED OF MODERNIZING

The system of granting tenure to teachers in American K-12 public schools began in the early 1900s as an effort to protect teachers from unfair and discriminatory dismissal practices, which were common at the time. Before tenure laws existed, teachers had little or no protection and could be summarily dismissed for "... speaking up, questioning educational practices, or merely because an administrator wished to give the job to someone else for political reasons or nepotism" (Baratz-Snowden). Today, every state but one— Wisconsin—mandates that school districts award teachers some form of tenure.

Tenure is generally awarded to teachers after passing a brief probationary period, which then affords them due process protections that are specified in state tenure law and the local collective bargaining contract.

Many people incorrectly believe that tenure means a teacher cannot be terminated. In fact, tenure does not mean teachers cannot be fired, but because of the cumbersome, lengthy, and costly due process protections it affords, many school districts rarely attempt to fire teachers—in effect granting them permanent employment.

A Job for Life. Although tenure technically does not require continuing the employment of an incompetent teacher (all tenure laws provide for dismissal of incompetent or inefficient teachers), in practice very few teachers are dismissed for any reason other than egregious conduct violations. Only 2.1 percent of all teachers are dismissed for poor performance annually, meaning that tenured teachers in most states enjoy a "job for life," regardless of their performance in the classroom (McGuinn).

What was originally designed to protect the nation's school teachers during an era of partisan school boards, corruption, and cronyism has now evolved into a system that automatically secures a lifetime position for teachers, regardless of their impact on students or the broader school community. As one critic observed, "In short, most school districts grant tenure not on the presence of recognizable achievement, but on the absence of criminal behavior" (Greenwald).

ABOUT THE TE² COMMUNITY BRIEFS

Community: A group of people with a common characteristic or interest living together within a larger society (Merriam-Webster Dictionary).

Communities for Teaching Excellence believes that every community and all of its members, including teachers, parents, community-based organizations, and interested citizens, benefit from inclusive and meaningful engagement on education issues.

Guided by this fundamental principle, we have created a series of documents called TE² Community Briefs. This series consists of several brief, informative, and research-based pieces on a variety of teaching effectiveness and equity (TE²) topics, such as fair evaluation, seniority, strategic compensation, and professional development, among others.

After an extensive review of the work being done around TE², we determined that a majority of the work tends to be either lengthy, academic research reports, or short op-eds and blogs. Few, if any documents are intended to educate and engage the community. The TE² Community Briefs fill this void by providing the reader with a comprehensive, research-supported summary of select TE² topics, including examples of states and school districts that are doing this work.

This brief discusses the shortcomings of existing tenure systems and promising approaches to fixing them, and includes an examination of two states that are leading the way in tenure reform.

Hindering Achievement. Current tenure practices do little to ensure that teachers are helping students achieve. Of the 49 states that mandate teacher tenure, only eleven—Colorado, Delaware, Illinois, Indiana, Massachusetts, Michigan, Nevada, New York, Oklahoma, Rhode Island, and Tennessee—require districts to incorporate minimal evidence of teaching effectiveness or general job performance into tenure decisions. The 38 remaining states permit school districts to award tenure virtually automatically (National Council on Teacher Quality, 2011a).

Current tenure practices do little to ensure that teachers are helping students achieve.

According to *The Widget Effect*, more than 40 percent of administrators reported they had never failed to renew a probationary teacher for performance concerns in his or her final probationary year (Weisberg, et al.). This is a missed opportunity because it is the last chance to dismiss low performing teachers before granting them tenure. A report from the Brookings Institution concluded, "Schools could substantially increase student achievement by denying tenure to the least effective teachers" (Gordon, et al.).

Snap Judgments. Thirty-eight states allow teachers to earn tenure in three years or less, which is not enough time for schools to accumulate the necessary evidence about a teacher's performance (National Council on Teacher Quality, 2011a). Typically, supervisors are given only two years to assess a newly hired teacher's instructional quality and to predict whether he or she will continue to develop. Not only are these brief probationary periods inadequate to judge who belongs in the teaching profession, they are also insufficient to grow and nurture teacher talent.

It is also important to note that in practice, the time frame during which school administrators must make decisions about granting tenure to probationary teachers is actually much shorter than the period specified by law, due to the logistics of recruitment, hiring, and staffing. For example, in California, state law mandates that teachers be granted tenure after two years of teaching, and that they must be notified by March 15th if they are to be dismissed. An analysis by the National Council on Teacher Quality (2011b) found that, "California policy results in districts using fewer than two years of information —and possibly only one formal evaluation— to assess a teacher's candidacy for tenure."

As a consequence, supervisors are often forced to make "snap judgments" about the quality of new teachers (Sutton).

COMMUNITIES for TEACHING EXCELLENCE

GETTING STARTED

Proposals to reform teacher tenure are starting to gain traction. However, tenure reform is most effective when it is paired with a comprehensive system for evaluating teaching performance, which includes defining and measuring teaching effectiveness. (For more information on teacher evaluation see: *Making it Meaningful: Building a Fair Evaluation System,* September 2011, and *Teaching Effectiveness: The Beginning of a Movement,* July 2011, Communities for Teaching Excellence.) As a report by the Center for American Progress asserts, "If tenure is to meet its twin goals of identifying and retaining an effective work force on the one hand, and weeding out weak and incompetent teaching, then it must be based on a strong, comprehensive evaluation system specifically designed to support best practice and build in due process … " (Baratz-Snowden).

Efforts by states such as Oregon, Alabama, Idaho, Mississippi, Texas, and Utah to reform tenure without first putting in place an evaluation system have been described by some as "futile," because even though these states have "replaced" tenure with renewable contracts (legal contracts that specify the period of work and terms, and that must be renewed in order for a teacher to work), in practice these contracts are virtually always renewed (McGuinn).

As part of the effort to reform tenure policies, states and school districts must work towards developing a comprehensive definition of teaching effectiveness and adopt unbiased, research-based methods for determining which teachers have met performance standards. These evaluation systems must provide feedback that best identifies the necessary support and training for teachers. Teachers should be supported early and often throughout their careers. Without such a system, it is very difficult to identify ineffective teachers, let alone justify dismissing them (McGuinn).

MORE THAN A RUBBER STAMP: CRITICAL ELEMENTS OF TENURE REFORM

Historically, elected officials have shown an overall reluctance to revise tenure laws. However, in recent years, several governors and even President Obama have called for tenure reform. An increasing number of states—Delaware, Florida, Illinois, Michigan, Ohio, Rhode Island, and Tennessee—are undertaking the challenging process of changing their tenure laws. While some states have proposed eliminating tenure altogether, others are attempting to strike a balance between holding public school teachers accountable for student outcomes, while still affording some measure of job protection and due process.

In recent years, several governors and even President Obama have called for tenure reform.

As the National Council on Teacher Quality (2010) notes, the awarding of tenure should be "more than just a rubber stamp." Rather, the process for determining who earns tenure should be "… a real evaluation of teacher quality and a deliberate decision about whether a probationary teacher should be granted this status—and the additional due process rights tenure brings—in a school system."

As the National Council on Teacher Quality notes, the awarding of tenure should be "more than just a rubber stamp."

In general, there are two primary aspects to tenure reform: raising the bar for the receipt of tenure and revising the dismissal process once a teacher has received tenure.

The first ensures that only the most effective teachers are awarded tenure. The second ensures that tenured teachers receive a fair and adequate hearing, while removing the consuming and expensive hurdles that make the dismissal of chronically ineffective, tenured teachers almost impossible.

Researchers and policymakers agree that meaningful tenure reform, which serves both teachers and students, should incorporate the following key elements and processes.

Lengthening Probationary Periods.

A majority of studies indicate that a teacher's effectiveness tends to be established by the fourth year of teaching, with effective teachers remaining relatively effective and ineffective teachers remaining relatively ineffective (Boyd, et al.). Yet, more than 80% of states currently grant tenure after three years or less, and only 12 states have probationary periods longer than three years. This is not enough time to allow teachers to demonstrate their ability to be effective at helping students learn.

Based on the research, tenure should not be granted before a teacher has been in the classroom for four years. For many tenure reform advocates, the ideal probationary period would be five years (National Council on Teacher Quality, 2010)

TENNESSEE AND MEMPHIS CITY SCHOOLS: TYING TENURE TO STUDENT SUCCESS

In April 2011, Republican Governor Bill Haslam signed groundbreaking legislation to significantly reform teacher tenure. The legislation extends the probationary period for new teachers from three years to five and ties tenure decisions—including maintaining tenure status—directly to the state's new teacher evaluation standards, which mandate 50% of a teacher's evaluation be based on student academic growth.

The law further requires probationary teachers to place in the top two tiers ("above expectations" or "significantly above expectations") of a new five-tier evaluation system in both the fourth and fifth years of teaching to receive tenure. It also allows teachers who receive tenure following the enactment of the law to be returned to probationary status if they rank in the bottom two tiers ("below expectations" or "significantly below expectations") for two consecutive years.

Teachers who received tenure prior to the enactment of the law cannot lose tenure status; however, the legislation expands the definition of "inefficiency"—a legal ground for dismissing current tenured teachers—to include being evaluated as "below expectations" or "significantly below expectations" (Locker).

Increasing the probationary period to receive tenure and connecting the decision to the state's new evaluation system are designed to improve student achievement statewide, since Tennessee has consistently ranked in the bottom quartile nationally (State Collaborative on Reforming Education). It also provides schools with more time to evaluate teachers and provide them with professional development before making tenure decisions.

Memphis City Schools (MCS) Superintendent Kriner Cash is a supporter of tenure reform because "it allows teachers more time in the field to perfect their skills" (Roberts). As Tennessee's largest school district MCS is proving to be a leader in evaluating and supporting probationary teachers through its Teacher Effectiveness Initiative (TEI), launched in 2009. The initiative's evaluation system, known as the Teacher Effectiveness Measure (TEM), is being used to inform tenure and other decisions. In June 2011, the Tennessee State Department of Education unanimously approved the use of TEM by districts throughout the state, recognizing the vital role of TEM in making tenure meaningful and increasing student achievement.

Tying Tenure to Performance.
The granting of tenure should be tied to demonstrated teaching effectiveness and evidence of student learning. Further, once a teacher has received tenure, he or she should be required to demonstrate continued effectiveness.

Many teachers agree that the current system of awarding tenure does not ensure that teachers are competent, and that probationary teachers should be required to demonstrate effectiveness. In a 2008 national survey of more than 1,000 teachers conducted by Education Sector, 69% said that when they learn that a teacher at their school has been awarded tenure, they think that it's "... just a formality—it has little to do with whether a teacher is good or not." Nearly 80% of teachers in the survey supported strengthening the formal evaluation of probationary teachers "... so that they will get tenure only after they've proven to be very good at what they do." And a majority (57%) think that even tenured teachers should be formally evaluated on a regular basis (Duffet, et al.).

Research confirms that allowing only effective teachers to earn tenure would lead to substantial increases in student achievement (Gordon, et al.). Similarly, the critical relationship between teacher quality and student achievement is well established, and states and districts should put in place tenure and evaluation policies to ensure that all students have access to effective teaching.

COLORADO: RAISING THE BAR FOR EARNING AND KEEPING TENURE

In 2010, Colorado passed the Great Teachers and Leaders Bill, tying tenure to student performance. The state is working to close the achievement gap and to increase academic performance, especially for low income students and students of color. While 79% of White students in 2010-11 scored proficient or above on Colorado's state reading assessment, the average for Black and Latino students was 49%. In math, 36% of Black and Latino students scored proficient or above on the state math assessment compared to 66% of White students. The achievement gap between economically disadvantaged students and their non-economically disadvantaged counterparts is similarly broad: 49% of economically disadvantaged students were proficient or above in reading, and 40% were proficient or above in math, whereas the numbers for non-economically disadvantaged students were 80% in reading and 67% in math (Colorado Department of Education).

Colorado passed its tenure reform provisions in May 2010 under the leadership of a Democratically-controlled legislature, a Democratic governor, Bill Ritter, and with support from the American Federation of Teachers (AFT)-Colorado. Beginning in the 2013-14 school year, new teachers must complete three consecutive years of teaching with evaluation ratings of "effective" or better in order to earn tenure. The law further requires that tenured teachers demonstrate effectiveness or face losing tenure and possible dismissal. Tenured teachers who receive two consecutive "ineffective" evaluations will lose their tenure, but will be offered a remediation plan. The law also provides for a "fair and transparent" appeal process to be "developed, where applicable, through collective bargaining." In an effort to streamline the process, the law stipulates that the appeal process cannot exceed 90 days (Colorado Senate Bill 191).

Essential to the implementation of these reforms is the development of a new performance evaluation system. The new system will be piloted in 2012-13 and implemented statewide in 2013-14.

Refining Dismissal Processes. Nearly half of the 1,000 teachers who participated in a recent national survey said that they personally know a tenured teacher who is ineffective and should not be in the classroom; and more than half (55%) indicated that in their district it is very difficult and time consuming to remove clearly ineffective teachers (Duffet, et al.). But on a more positive note, in an earlier survey, nearly 2 to 1—57 percent to 29 percent—believed that it is possible to change tenure rules and the discipline process in a way that permits poor-quality teachers to be dismissed more easily and still protect job security rights (Henke, et al.).

So, how might states and districts put in place dismissal procedures that are less cumbersome but still ensure fairness and due process? As described above, the starting point is to develop a transparent and comprehensive evaluation system designed to support best practices and help teachers improve.

States and districts should also take measures to ensure that those who are in the position of making judgments about teacher effectiveness—and ultimately about tenure and dismissal—are trained professionals who understand teaching and learning. A report from the Center for American Progress argues that professional educators make decisions concerning teacher quality and competence rather than administrative law judges, and that the process be developmental rather than adversarial (Baratz-Snowden).

Refining current dismissal policies presents an opportunity for collaboration between teachers, unions, and administrators.

While a number of policymakers believe that present day civil rights and labor laws protect teachers from unfair dismissal, many teachers strongly believe that they need additional protections from incompetent or vindictive administrators, and overzealous parents. However, legislation such as that passed in Colorado shows that it is possible to reform tenure, advance teaching effectiveness, refine dismissal processes, all while supporting teachers.

Finally, refining current dismissal policies presents an opportunity for collaboration between teachers, unions, and administrators. Some school districts, such as those in Minneapolis, Minnesota and Toledo, Ohio, have partnered with unions to develop successful systems—based on peer assistance and review models—for holding teachers to high standards in earning tenure and for protecting due process in dismissal actions.

One researcher who studied the Toledo model, which has been in place since 1981, found it to be more rigorous than traditional methods for granting tenure review and for terminating weak teachers (Kerchner, et al.).

Teachers' unions, school districts, and states alike have much to gain from modernizing present day tenure systems. By working together to develop clear standards of excellent practice, the tools and procedures to measure that practice, and rigorous, fair, and streamlined systems for removing teachers who are not meeting the standards, they can make certain that all students receive the quality education they need and deserve.

Teachers' unions, school districts, and states alike have much to gain from modernizing present day tenure systems.

WHAT CAN YOU DO?

Find out about the teacher tenure system in your state. How long is the probationary period before tenure can be granted? Is the awarding of tenure tied to performance? Get informed about local efforts to reform tenure and dismissal processes and to improve teaching effectiveness. For more information, or to get involved in the TE² movement, contact Communities for Teaching Excellence at www.4teachingexcellence.org, and be sure to check out the exciting work going on in Hillsborough County, FL; Memphis, TN; Pittsburgh, PA; and The College-Ready Promise in Los Angeles, CA. Together, we're working to ensure effective teaching for every student, in every classroom, every year.

COMMUNITIES for TEACHING EXCELLENCE

SOURCES & CREDITS

Baratz-Snowden, J. (2009). *Fixing Tenure: A Proposal for Assuring Teacher Effectiveness and Due Process.* (Washington, D.C.: Center for American Progress), pp. 1, 10.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J. and Wyckoff, J. (2007). *The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools.* (Washington, D.C.: Urban Institute).

Colorado Department of Education (2011). *State Performance Data.* Retrieved from http://www.schoolview.org/performance.asp

Colorado Senate Bill 191 (2010). *Concerning Ensuring Quality Instruction Through Educator Effectiveness.* Retrieved from http://www.leg.state.co.us/CLICS/CLICS2010A/csl.nsf/fsbillcont3/EF2EBB67D47342CF872576A80027B078?Open&file=191_01.pdf

Duffett, A. et al. (2008). *Waiting to be Won Over: Teachers Speak on the Profession, Unions, and Reform.* (Washington, D.C.: Education Sector Reports), p. 2.

Frey, D. (2010). *State Tenure/Continuing Contract Laws.* (Denver, Co: Education Commission of the States). Retrieved from http://www.ecs.org/clearinghouse/88/28/8828.pdf

Greenwald, R. (2010, August 24). A Modest Proposal for Tenure Reform. *In These Times.* Retrieved from http://www.inthesetimes.com/article/6325/a_modest_proposal_for_teacher_tenure_reform/

Gordon, R., Kane, T.J. and Staiger, D.O. (2006). *Identifying Effective Teachers Using Performance on the Job.* Hamilton Project Discussion Paper. (Washington, D.C.: The Brookings Institution), p. 13.

Henke, R., Choy, S. Chen, X., Geis, S., Alt, M., Broughman, S. (1997). *America's Teachers: Profile of a Profession, 1993–94.* (Washington, D.C.: Department of Education, National Center for Education Statistics, NCES 97-460).

Kerchner, C.T., Koppich, J.E., and Weeres, J.G. (1997). *United Mind Workers: Unions and Teaching in the Knowledge Society.* (San Francisco: Jossey-Bass).

Locker, R. (2011, April 12). Tennessee Gov. Haslam Signs Teacher Tenure Bill into Law. *The Commercial Appeal.* Retrieved from http://www.commercialappeal.com/news/2011/apr/12/gov-haslam-signs-tenure-bill-law/

McGuinn, P. (2010). *Ringing the Bell for K-12 Tenure Reform.* (Washington, D.C.: Center for American Progress).

National Council on Teacher Quality (2011a). *State Teacher Policy Yearbook: National Summary.* (Washington, D.C.).

National Council on Teacher Quality (2011b). *Teacher Quality Roadmap: Improving Policies and Practices in LAUSD.* (Washington, D.C.), p. 36.

National Council on Teacher Quality (2010). *Blueprint for Change: A National Summary.* (Washington, D.C.), p. 10.

Rivkin, S. G., Hanushek, E. A. and Kain, J. F. (2005). "Teachers, Schools, and Academic Achievement." *Econometrica,* Vol. 73, No. 2, 417– 458.

Roberts, J. (2011, March 24). Governor's Tenure Reform Bill for Teachers Passes House. *The Commercial Appeal.* Retrieved from http://www.commercialappeal.com/news/2011/mar/24/governors-tenure-reform-bill-teachers-passes-house

Sawchuk, S. (2010, April 7). States Strive to Overhaul Teacher Tenure. *Education Week,* pp. 1, 18.

State Collaborative on Reforming Education (2011). *The State of Education in Tennessee-2010.* (Nashville, TN).

Sutton, P. (2009, December 16). Thinking Anew About Teacher Tenure. *Education Week Commentary,* pp. 20-21.

Tennessee Department of Education (2011). *New Tenure Law: Frequently Asked Questions.* Retrieved from http://www.tn.gov/education/doc/NewTenureLawFAQs4.27.11.pdf

OUR MISSION

Communities for Teaching Excellence works to improve students' academic achievement and their futures by empowering communities to advocate for effective teaching for every student, in every classroom, every year.

ACKNOWLEDGEMENTS

Many thanks to Kate Tromble for her comments, edits, and feedback, and her overall willingness to improve the quality of the publication.

Communities for Teaching Excellence

A project of Rockefeller Philanthropy Advisors
448 South Hill St., Suite 408
Los Angeles, CA 90013 | 213.489.3002
4teachingexcellence.org

EXHIBIT 12
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

New York Law Journal

WWW.NYLJ.COM

An ALM Publication

VOLUME 249—NO. 94

THURSDAY, MAY 16, 2013

Outside Counsel

Expert Analysis

Evaluation Law Could Limit Ability To Terminate Probationary Teachers

By
Warren H. Richmond III

It has been over a year since Governor Andrew Cuomo announced an "historic" settlement between the New York State United Teachers and the State Education Department which he predicted would make New York State "a national leader in holding teachers accountable for student achievement." The statute, Education Law Section 3012-c and its implementing Regulations, 8 NYCRR Subpart 30-2, which were the product of this settlement, are collectively known as APPR (Annual Professional Performance Review). Together they create a comprehensive and complex evaluation system for rating teachers and principals which places strong emphasis on student achievement and growth as reflected on standardized tests.

As part of this system, teachers are given a numerical score which is then transposed into a rating of "highly effective," "effective," "developing" or "ineffective." Ironically, while the intent of the APPR initiative is to improve teacher performance, another consequence of the legislation is that it will be significantly more difficult for school districts to terminate non-tenured teachers whose performance is inadequate or otherwise problematic.

Prior to the enactment of Education Law Section 3012-c, school districts possessed broad discretion to terminate teachers prior to their being granted tenure. Indeed, more than 37 years ago in *James v. Board of Education of Central School District No. 1 of the Town of Orangetown and Clarkstown*, 37 NY2d 891, 892 (1975) the Court of Appeals stated:

> A board of education has an unfettered right to terminate the employment of a teacher during his probationary period unless the teacher establishes that the board terminated for a constitutionally impermissible purpose or in violation of statutory proscription.

In the years since the *James* decision, the principle enunciated in that case has, with limited exception, governed the termination of probationary teachers in New York State. See e.g. *Conetta v. Board of Ed. of Patchogue Medford UFSD*, 165 Misc.2d 329 (Sup. Ct. Suffolk Co. 1995) (tenure cannot be denied on the basis of the board's philosophical opposition to tenure). Although Education Law Section 3031 provides a procedure by which a Superintendent of Schools is required to set forth his or her reasons for recommending termination or a denial of tenure, the courts have held that this process is designed only to allow probationary teachers to ascertain whether any of the reasons were constitutionally or statutorily impermissible. It is not meant in any way to restrict the discretion afforded the Superintendent and the Board of Education. See *Merhige v. Copiague School District*, 76 AD2d 926 (2d Dept. 1980).

The enactment of Education Law Section 3012-c has substantially expanded the protection given to probationary teachers. Section 3012-c(1) specifically provides:

> [A]nnual professional performance reviews shall be a significant factor for employment decisions including but not limited to promotion, retention, tenure determination, termination, and supplemental compensation, which decisions are to be made in accordance with locally developed procedures negotiated pursuant to the requirements of article fourteen of the civil service law where applicable. Provided, however, that nothing in this section shall be construed to affect the statutory right of a school district or board of cooperative educational services to terminate a probationary teacher…for statutorily and constitutionally permissible reasons other than the performance of the teacher…in the classroom…, including but not limited to misconduct.

There can be little question that the above language modifies the long-established rule that a board of education possesses the "unfettered right" to terminate a probationary teacher absent reasons that are constitutionally impermissible or in violation of a statute. What remains unclear, however, is the extent to which this has occurred. Much, but by no means all, of the problem results from the failure of the drafters of Section 3012-c to define two of the pivotal terms in the statute, i.e. "significant factor" and "performance." The meaning and application of these terms, which will ultimately be left to the courts, will to a great extent set the parameters of the discretion afforded to school boards in making the important decisions as to which members of the teaching staff will obtain tenure.

A consequence of the legislation is that it will be significantly more difficult for school districts to terminate non-tenured teachers whose performance is inadequate or otherwise problematic.

Significant Factor

In providing that the APPR review will be a "significant factor" in employment decisions including tenure determination, Education Law Section 3012-c provides little concrete guidance. It is clear that at a minimum the APPR must be considered in making such decisions. On the other hand, the statute falls short of requiring, as it easily could have, that the APPR review be the determining factor. Thus, the extent to which the APPR rating is to be considered is likely somewhere in between. The problem for school districts, and for that matter for teachers,

WARREN H. RICHMOND III *is a partner at Harris Beach in the Long Island office.*

is that neither the statute nor its implementing regulations provide any guidance whatsoever as to the nature of the other factors that may be taken into consideration to outweigh an APPR rating of "effective" or "highly effective."

For example, to what extent may a school district consider, and what weight may be accorded, more subjective factors such as ability to get along with other staff, ability to communicate with parents, or concerns about poor judgment? Factors such as these, which are not readily quantifiable, were entirely appropriate considerations in tenure determinations prior to the enactment of Section 3012-c. Indeed, denial of tenure on the basis of considerations of this nature was essentially unreviewable.

While there is nothing in Section 3012-c to suggest that these more subjective considerations are now precluded, it appears that the extent to which they can be the basis for the denial of tenure will necessarily be subject to review. Specifically, a court may be called upon to determine whether such a consideration outweighed the "significant factor" of an "effective" or "highly effective" APPR rating. In this light it is not difficult to see that in many cases it will be the courts, not the board of education, that make the ultimate determination regarding teacher termination.

Performance

Section 3012-c carves out an exception to the consideration of APPR ratings in the making of employment decisions stating,

> …nothing in this section shall be construed to affect the statutory right of a school district to terminate a probationary teacher… for statutorily and constitutionally permissible reasons other than the performance of the teacher…in the classroom…including but not limited to misconduct.

However, the term "performance," like the term "significant factor," has been left undefined by the legislation's drafters. Was it their intent that the term be narrowly defined so as to refer solely to performance as reflected by the completed APPR score received by a teacher?

If that were the intended meaning a board of education would retain much of its discretion to determine whether or not to dismiss a probationary teacher. It would, for example, be able to terminate a teacher for largely subjective reasons such as concerns about poor judgment, notwithstanding an "effective" or even "highly effective" APPR rating. At the other end of the spectrum, "performance" in the classroom might be defined to mean anything that is related to teaching performance in its most general sense. Were that to be the meaning, a board's discretion would be significantly constrained. A board might well be precluded from terminating a probationary teacher for performance-related issues (e.g., classroom management or inadequate lesson planning) notwithstanding an "effective" or "highly effective" APPR rating.

Ultimately, the meaning of the terms "significant factor" and "performance" in the classroom will be defined through litigation. However, until such time as the courts or the Commissioner of Education provide direction regarding these key terms, school districts will remain very much in the dark as to the degree to which they possess discretion to terminate probationary teachers. Unfortunately, in an attempt to avoid litigation, some districts may err on the side of caution and grant tenure to teachers despite significant reservations as to their competence.

The term 'performance,' like the term 'significant factor,' has been left undefined by the legislation's drafters. Was it their intent that the term be narrowly defined so as to refer solely to performance as reflected by the completed APPR score received by a teacher?

Education Law Section 3012-c presents various other difficulties by school districts related to the employment of probationary teachers. First, because the APPR process will not be completed until the end of a school year at the earliest, a question exists as to the ability of a school district to terminate a probationary teacher during his or her first year of teaching. Can a district, for example, terminate a new teacher who has proven to be utterly ineffective after three or four months of teaching or must it allow such a teacher to continue in a classroom for the entire school year?

Guidance issued by the State Education Department is far from helpful, stating cryptically: "Prior to completion of the APPR in the first year of the probationary term, a probationary teacher…may be summarily dismissed for constitutionally and statutorily permissible reasons other than classroom performance without regard to the APPR." Guidance on New York State's Annual Professional Performance Review for Teachers and Principals to Implement Education Law §3012-c and the Commissioner's Regulations, Updated Aug. 13, 2012 (C-13 at p. 24)

Second, the timelines of the APPR do not align with the statutory timelines for decisions regarding teacher termination. The provisions of Education Law set forth a 60-day period in which to terminate a probationary teacher. The teacher is first entitled to 30 days' notice of the meeting at which the board of education will consider termination. Section 3031(a) and (b). Once the board has voted to terminate, the teacher is terminated on 30 days' notice. Section 3019(a).

As a result of these statutory notice periods, board action to terminate probationary teachers has generally taken place during the months of April and May so that the termination may be effective at the end of the school year. Now that Education Law Section 3012-c(1) requires that the annual professional performance review be a "significant factor" in the decision to terminate a probationary teacher, such action will likely be delayed.

Due to the necessity of incorporating end-of-year student achievement scores, the final APPR rating may not be provided until as late as Sept. 1 of the following school year. Section 3012-c(2)(c)(2). As such, the school district will effectively be precluded from terminating a probationary teacher at the conclusion of a year of poor performance. Given the statutory time periods contained in Section 3031(a) and (b) and 3019(a), such termination may not take place until well into the fall of the next year. Such delay may be further extended by virtue of the teacher taking an appeal from his or her APPR rating. Section 3012-c(5)(6).

Finally, neither the APPR statute, the Commissioners' Regulations nor the Guidance promulgated by the State Education Department provides any guidance as to what is to occur at the end of a teacher's probationary term. Pursuant to Education Law 3013, prior to the expiration of a teacher's probationary term, the superintendent of schools is required to make a recommendation to the board of education as to whether the teacher is to be granted tenure.

Section 3012-c(2) specifically requires that every person who is not to be recommended for tenure be notified in writing not later than 60 days immediately preceding the expiration of the probationary period. Because, in the vast majority of cases, the end of the teacher's probationary term corresponds with the end of the school year, it will effectively be impossible to include the final year's APPR as a "significant factor" in the tenure determination as required by Section 3012-c(2). Adding to this difficulty is the fact that districts will not have the luxury of continuing the teacher's employment into the following school year as it will result in tenure by estoppel. See, e.g., *Lindsey v. Board of Education of Mt. Morris Central School District*, 72 AD2d 185 (4th Dept. 1980).

The issues raised above and no doubt many others related to the application of APPR to probationary teachers will be subject to much litigation in the coming years. Given the significance of tenure, which in effect represents a lifetime job, it is hoped that in applying the provisions of Section 3012-c the courts will, to the greatest extent possible, preserve the sound discretion of school administration to retain only those who they believe are capable of providing quality education.

Reprinted with permission from the May 16, 2013 edition of the NEW YORK LAW JOURNAL © 2013 ALM Media Properties, LLC. All rights reserved. Further duplication without permission is prohibited. For information, contact 877-257-3382 or reprints@alm.com. #070-05-13-32

EXHIBIT 13
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

How Do I Fire an Incompetent Teacher?

You are the administrator of a New York City public school. Your mission is to fire a teacher who is so inept that virtually no other teacher, parent, or student would object.

Problem: The teacher has tenure.

In New York City, teachers become eligible for tenure after three years. State law grants teachers tenure after five years of continuous service.

STEP ONE: Evaluate and document.

Documentation requires placing notes in the teacher's file to record incompetence.

Teachers cannot be disciplined for "micromanagement" violations, meaning bulletin board format, furniture arrangement, or durations of lesson units.

Every note must be shown to the teacher, and she must sign the note. The teacher's signature means only that she has seen the note, not that she agrees with its content.

The Department of Education will be issuing "detailed guidance on best practices for issuing letters to file shortly."

Notes may not be added more than three months after an incident.

The teacher can attach a written response to any letter in the file.

Letters that do not result in disciplinary charges must be removed from the teacher's file after three years.

The Department of Education will issue additional guidance on the removal of file letters shortly.

You may proceed past this point only if enough documentation has accumulated in the teacher's file to justify your giving the teacher an unsatisfactory rating on the teacher's annual evaluation.

You think you have enough evidence and assign an unsatisfactory rating.

You're ready for a round of appeals.

Good news! The burden from here on out is largely on your school's lawyer. You can watch while others decide what staffing decisions are best for your school.

Don't give up.

You're ready to fire this teacher.

The teacher first appeals to a committee designated by the chancellor.

At the hearing before the committee, the teacher may be advised by an employee of the City School District or a union representative.

The teacher will be confronted by witnesses and may call witnesses, examine exhibits, and introduce evidence.

The committee renders a decision.

The committee overturns the rating. Return to STEP ONE.

The committee upholds the rating.

But its decision isn't final. The committee forwards its decision to the chancellor for a "final" decision.

The chancellor overturns the rating.

Return to STEP ONE.

The chancellor upholds the rating.

The aggrieved has 30 days after the action to appeal.

Section 310 of the New York State Education Law provides that anyone who considers herself aggrieved by an action taken by school authorities may appeal to the commissioner for a review.

Appeal to the state commissioner of education.

Option 1

But to protect the teacher from arbitrary or unfair dismissal, she has a few appeal options.

Option 2

File in state court to have the chancellor's decision overturned.

Option 3

Don't appeal.

The State Department of Education says, "We endeavor to issue a decision within six to eight months (i.e. eight to ten months after the petition is filed). However, hundreds of complex appeals are filed each year, so it is impossible to guarantee a specific date."

So the inept teacher will be around a while longer. But at least the commissioner gives "great deference to decisions by local officials" and will not substitute his judgment unless "he can find no rational, reasonable basis for the district's actions."

Convinced? Here are some actual decisions:

The commissioner decided that a teacher should have been suspended for a year, rather than dismissed, because even though she was guilty of "excessive absences which had a detrimental impact on students...there was no indication that teacher was guilty of malingering, and the record indicated that she had tried (to some extent) to reduce the number of her absences and to assist some substitute teachers hired to replace her." 1994 Op Comm Ed No. 13278.

The termination of a licensed probationary special education teacher was not upheld because the board didn't adhere to a circular issued by the chancellor regarding the tenure and probation of teachers. The circular said that a special education teacher, before being terminated, was entitled to the "substantial rights of an evaluation by the Division of Special Education and Pupil Personnel Services." *Lehman v. Board of Education* (1981, 2d Dept) 82 AD2d 832, 439 NYSd 670.

To the courtroom.

Even if the teacher doesn't pursue another appeal at this stage, she cannot be fired yet. Termination can result only from a hearing under Section 3020-A of the New York State Education Law.

File charges to terminate the inept teacher.

Just to recap, the inept teacher has *not* been fired yet. You have succeeded at placing notes in her file and assigning an unsatisfactory rating. You may now proceed to the termination hearings under Section 3020-A of the New York State Education Law.

The commissioner finds that the teacher was fairly assigned an unsatisfactory rating and that the rating was appropriately upheld in a committee proceeding adhering fully to established procedures.

Back to STEP ONE.

Continued next page

How Do I Fire an Incompetent Teacher? part 2

Continued from previous page

After the charges are filed, the board, in executive session, will vote on whether the charges are potentially valid.

Not Valid? Obtain further documentation. Back to STEP ONE.

***Potentially* Valid.**

The School Board must issue a statement of the charges and maximum possible penalty and send a notice outlining the teacher's rights under Section 3020-A, including the right to a formal hearing.

Within 10 days, the teacher can request a hearing and elect to appear before either a single hearing officer or a panel of arbitrators. In any event, a hearing officer will run the hearing.

If a teacher is to be suspended pending the outcome of the hearing, it will be a *paid* suspension unless a "probable cause arbitrator" finds there is probable cause that the charges involve one of the following:

The felony sale, possession, or use of marijuana, a controlled substance, or a precursor of a controlled substance or drug paraphernalia; or

Any crime involving physical or sexual abuse of a minor or student; or

Any felony committed either on school property or while in the performance of teaching duties, or any felony involving firearms (other felonies may not be sufficient).

If the teacher is convicted of or is pleading guilty to a felony, she will be suspended without pay for no more than three months pending the outcome of the hearing.

Before the Hearing

Both sides will exchange witness lists, witness statements, and physical evidence (e.g., photographs), at least before the presentation of their direct case.

The school board must provide the teacher with copies of investigatory statements, notes, other exculpatory evidence, and relevant student records (after the hearing officer reviews them first).

But the teacher has to give the school board only documents the board can prove are relevant.

Pre-Hearing Conference

Once the panel or hearing officer is in place, within 10 to 15 days, the hearing officer will hold a pre-hearing conference.

At the pre-hearing conference, the hearing officer can:

- Issue subpoenas;
- Hear and decide all motions, including motions to dismiss the charges;
- Decide any discovery disputes about materials one side or the other wants to have before the hearing.

Set the timetable for the hearing.

NOTE: The board can always decide to offer an expedited hearing to the teacher under any circumstances, but the teacher does not have to accept it.

There will be an expedited hearing.

What if the school board shows at the conference that the teacher's license has been revoked (and all appeals exhausted)?

Hearing Procedures:

The teacher shall have a reasonable opportunity to defend herself and an opportunity to testify.

But the teacher does not have to testify.

Each party, of course, has the right to have lawyers;

To subpoena witnesses;

And to cross-examine witnesses.

All testimony will be taken under oath.

A stenographer will transcribe the proceedings.

What happens after the hearing?

The hearing officer must give a written decision within 30 days of the last day of the final hearing.

Copies go to the commissioner of education.

The commissioner forwards copies to the teacher and to the clerk or secretary of the board.

In determining the penalty, the hearing officer must consider whether the school board made efforts to correct the behavior of the employee that resulted in charges. Did the board try remediation, peer intervention, or an employee assistance plan?

Yes.

If the hearing officer has ruled in the board's favor and found that the board made efforts to correct the employee's behavior, then you have finally succeeded in terminating one inept teacher.

No.

If not, the hearing officer can order efforts to correct the employee's behavior instead of dismissal.

Except that the teacher may appeal.

Within 10 days of the decision, either the teacher or the school board can appeal the decision to the New York State Supreme Court.

You are now involved in a lawsuit. Please consult your school attorneys for additional details.

EXHIBIT 14
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

3020-a process remains slow, costly

On Board Online • NYSSBA News • May 11, 2009

NYSSBA survey

By Patricia Gould
Assistant Counsel

School districts in New York State spend an average of $216,588 to pursue disciplinary charges against tenured teachers or administrators, according to a NYSSBA survey covering 2004 through 2008.

That's up from $128,941 in 2004 – an increase of 70 percent.

Even after adjusting for inflation, that's an increase of more than 49 percent.

The single most common type of charges involved improper sexual remarks, improper physical contact or improper relationships with students. Allegations involving insubordination and pedagogical incompetence were also common charges brought under the disciplinary statute, Section 3020-a of the state Education Law.

The survey was sent to all NYSSBA member districts and BOCES. Responses were received from 400 districts, a 59 percent response rate. The results do not include New York City, where disciplinary cases are handled through a negotiated process that is an alternative to 3020-a.

Other findings:

- Middle school and high school teachers were the personnel most frequently accused of misconduct under 3020-a (52 percent of reported cases). Elementary teachers were accused in 18 percent of cases while cases against administrators represented 6 percent.
- The largest expense of responding districts was the salary and fringe benefits paid to the suspended employee (52 percent of costs). The 400 responding districts spent a totaled of $7.4 million on salary and fringe benefits for individuals suspended while their 3020-a case was pending – a combined average cost of $136,676 per case.
- Paying salary and benefits to substitutes represented 30 percent of 3020-a costs, while legal expenses represented 12 percent of costs associated with 3020-a proceedings.
- Other expenses include other staff costs (5 percent) and miscellaneous costs, such as paying for outside investigators, expert witnesses, transcription, photocopying and travel (1 percent).

Lengthy process

It took an average of 502 days to conclude a full 3020-a hearing, from the date charges were levied to the date a decision was issued by a hearing officer or panel. Although this represents a drop of 18 days from the average of 520 days between 1997 and 2004, it still constitutes an alarming increase from the low of 319 days between 1994 and 1997.

The most time consuming part of the process occurred from the first to last day of hearing – an average of 176 days. In addition, there was an average of 136 days between the last hearing day and the date of a decision.

Forty-eight percent (48 percent) of districts responding to NYSSBA's survey considered bringing 3020-a charges at least once but did not. Their stated reasons for not filing charges were as follows:

- Employee resigned **49 percent**
- Employee retired **12 percent**
- 3020-a process too cumbersome **15 percent**
- 3020-a too expensive **17 percent**
- District case not strong enough **21 percent**
- Not enough documentation **15 percent**

When charges were filed, 46 percent were either settled or discontinued prior to a final decision. Hearing officers (or panels) issued a decision in only 18 percent of the cases reported in this survey. Thirty cases (36 percent) were pending at the time the survey was concluded in March 2009.

Penalties and appeals

In 37 cases reported, districts filed charges but reached a settlement with the employee without a hearing. In those cases, 20 teachers resigned or retired, six agreed to suspensions, five were fined, and three were terminated without a hearing because they did not demand one. Three other cases were resolved in other ways not further specified by the responding district.

In cases where a decision was rendered by a hearing officer, only three teachers were fully acquitted by the hearing officer. Among those found guilty of at least one charge, seven teachers were terminated, five were suspended without pay, two were reprimanded, and one fined.

Appeals were uncommon; only three reported cases were appealed to a court. Under the limited grounds for appeal prescribed by statute, a 3020-a hearing officer's decision can only be reversed on very narrow grounds and it is unusual for a court to reverse a decision under these standards. Thus, the decision of the hearing officer is likely to be final in most cases.

The state Legislature last revised the 3020-a law in 1994. When asked

Send this page to a friend

Show Other Stories

EXHIBIT 15
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

NYSSBA ISSUE BRIEF

3020-a Teacher Discipline Reform

Under current law a “3020-a” teacher disciplinary proceeding takes an average of 520 days from the date charges are brought to the date of a final decision; at an average cost of $128,000.00. Proceedings addressing pedagogical incompetence take an average of 830 days at an average cost of $313,000.00. The recent addition of an “expedited” process for those who receive two consecutive subpar evaluations is not nearly sufficient to address this issue. Real reform of the teacher discipline process is needed. Independent contractor arbitrators in disciplinary cases must be replaced by NYSED administrative law judges. Cases would be decided more quickly, enabling districts to return the teacher to the classroom or hire a permanent replacement. In either event, taxpayers would be relieved of paying for costly and needless delays. Many of the needed reforms just makes sense: For instance, teachers convicted of child abuse, those who have had their license to teach revoked and those who do not obtain permanent certification in the time required by law should be removed without onerous procedural requirements. Simply put, our state can no longer afford a process that is both ineffective and time consuming.

EXHIBIT 16
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

LOOK BELOW THE SURFACE FOR A SMARTER SOLUTION FOR THE HUDSON.
Learn More >
Safe. Secure. Vital.
Indian Point Energy Center

Firing teachers: mission impossible

Recent reforms aside, bad educators have a lock on their jobs

BY KATHARINE B. STEVENS / NEW YORK DAILY NEWS / Monday, February 17, 2014, 4:25 AM

A A A
81 31
SHARE THIS URL
nydn.us/1nzRSzi COPY

PAUL BUCKOWSKI/ALBANY TIMES UNION

NYSUT President Iannuzzi, UFT President Michael Mulgrew and Gov. Cuomo.

If there's one thing that's important to good schools, it's good teachers. That's the premise of a groundbreaking lawsuit now being heard in Los Angeles Supreme Court that challenges entrenched California state laws protecting the jobs of public school teachers who are "grossly ineffective."

The nine student plaintiffs, Vergara vs. California, argue that laws protecting even abysmally incompetent teachers violate the guaranteed civil right of the state's children to access a decent education.

The outcome of the case could have big implications for New York. Despite the state's highly visible new teacher evaluation law and a perception of radical change under the Bloomberg administration, a scandalous reality remains: Here, as in California, it is virtually impossible to dismiss a grossly ineffective teacher.

This is because the particularly crucial New York State law governing teacher dismissal, 3020-a, was left essentially untouched by both the contentious new teacher evaluation legislation and recent city reforms.

Under that law, only the state can dismiss a teacher — and it rarely does.

I recently completed an in-depth study of a decade of official reports on the state's dismissal hearings for New York City teachers. My analysis shows that the problem of extraordinary job protection for grossly ineffective teachers in New York is worse than many understand.

GET U READY
FROM CLASSROOM TO DORM ROOM
We Have Everything You Need for College
P.C. RICHARD & SON
Shop Now

EDITORS' PICKS

Charles not in charge
The City Council, led by Speaker Christine Quinn, did the right thing yesterday in stripping Brooklyn Councilman Charles Barron of the chairmanship of the

She's got the blues! Singer dyes her locks to colorful hue
Some are hitting the (dye) bottle, others are chopping it off. It's hard to keep up in Hollywood, where stars

Giants Insider: Reese hopes unknown TE catches on
Jerry Reese still believes there could be a star among the Giants' anonymous group of tight ends, and he

Rangers and Chris Kreider agree to a two-year, $4.95M deal
On second thought, let's make a deal. That's what the Rangers and Chris Kreider's agent decided

Joba goes 'Duck Dynasty' in Detroit: Hairdos and hair-don'ts in sports
The larger-than-life personalities in sports often try to match their attitudes

FROM AROUND THE WEB

15 Of the Hottest Female Sports Broadcasters
(RantSports)

Even attempting to get the state to dismiss a teacher is prohibitively expensive and burdensome. According to the New York State School Boards Association, the average 3020-a proceeding for a single incompetent teacher extends for 830 days and costs taxpayers $313,000.

Over the 10-year period I studied (1997-2007), just 12 of New York City teachers (of whom there are 75,000 at any given time) were dismissed for incompetent teaching. Teachers who had years of "unsatisfactory" ratings; who were proven over months of hearings to be grossly incompetent; who were verbally and physically abusive to children, parents and colleagues, or who simply failed to come to work for days and weeks on end were returned to classrooms.

My analysis further reveals that the minimum level of teaching effectiveness required for tenured teachers to keep their jobs in New York City schools is defined not by the schools (much less parents and communities) but behind closed doors in arbitration proceedings controlled by the state.

In practice, teachers are dismissed only if they are proved to have been grossly ineffective and "incorrigible," without even a remote possibility of improving. That is, the operative state standard for returning a teacher to the classroom is not demonstrated effectiveness, but a teacher's potential capacity to be even marginally competent in the future.

The new state evaluation law, currently wrapped up in debate over the Common Core standards, is supposed to change this by making it far easier to fire teachers who are rated "ineffective" two years in a row.

But in the first year of the new evaluation system, just 1% of teachers received that rating. And all dismissals will still go through 3020-a, which makes removal almost impossible.

The vast majority of New York City's teachers are responsible, effective and hardworking, and the problem of grossly ineffective teachers is often dismissed as just "a few bad apples." Yet if even just 1% of the city's teachers are ineffective, then 10,000 students in New York City have an ineffective teacher every day.

New York City's new chancellor, Carmen Fariña, knows that removing ineffective teachers is essential to creating good schools. While serving as principal of PS 6 on Manhattan's Upper East Side in the 1990s, she replaced 80% of the school's teaching staff — not by dismissing them, which is illegal under state law, but by persuading them to leave, usually for positions in other, less-demanding schools.

Ironically, as New York City's schools chancellor, she now has no formal power to replace even one teacher.

This a crucial educational equity issue. A child assigned to the classroom of a grossly ineffective teacher is denied the opportunity for the sound, basic education guaranteed by the New York constitution.

New York's laws should be ensuring that all children have effective teachers, not making it impossible to cut loose those who are incompetent, indifferent or even abusive.

Let's hope Vergara prevails in California. And let's overturn dysfunctional New York State laws that protect ineffective teachers' jobs at the expense of children and the public schools.

Stevens received a Ph.D. in education policy from Columbia University in 2013. She was formerly director of Teachers for Tomorrow, a program that prepares teachers to work in low-performing New York City schools.

PROMOTED STORIES

The Cop Who Arrested Me Was A Good Man
(New York Natives)

Gun-rights activist emailed a pro-gun safety writer. You'll easily believe what happened next.
(Blue Nation Review)

Rude! What You Should Never Do in Other Countries
(Reader's Digest)

Recommended by outbrain

EDITORS' PICKS

Ex-Chelsea star Frank Lampard officially signs with NYCFC
It's official. Frank Lampard is a member of New York City FC. The England international and long-time

Guess which star is soaking up the sun in this sexy snap?
Kim's usual flashy Hollywood party girl style has taken a strange turn since she fell for Kanye

Yankees Insider: Brian Cashman says Masahiro Tanaka feeling less pain
Masahiro Tanaka is starting to feel less pain in his elbow, although he's still far

Harper: Here is why Rockies All-Star shortstop Troy Tulowitzki won't be making the move to
Troy Tulowitzki may be the holy grail for Mets' and

Bartolo Colon flirts with perfect game, dominates Mariners in 3-2 Mets win
If it was his last start in a Mets uniform, at least it was memorable. While the Mets

Giants believe Eli Manning can still be Super Bowl level quarterback
Eli Manning is coming off a season in which he threw a career-worst 27

HOT MAMA! Jennifer Lopez defies age in skimpy bikini
She may be a mother of twins, but Jennifer Lopez is hotter than ever. Check out J. Lo's sexiest (and most

July 24: Traffic safety, Eric Garner and Elvis Presley
Flushing: Several months ago, I wrote to Voice of the

EXHIBIT 17
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

SUPERIOR COURT OF THE STATE OF CALIFORNIA

COUNTY OF LOS ANGELES

FILED
Superior Court of California
County of Los Angeles
JUN 10 2014
Sherri R. Carter, Executive Officer/Clerk
By K. Mason Deputy

BEATRIZ VERGARA, a minor, by Alicia Martinez, as her guardian ad litem, *et al*, Plaintiffs,	Case No.: BC484642
vs.	TENTATIVE DECISION
STATE OF CALIFORNIA, et al, Defendants	Dept. 58
CALIFORNIA TEACHERS ASSOCIATION, *et al*, Intervenors	Judge Rolf M. Treu

In accordance with California Rules of Court 3.1590, this Court now announces its Tentative Decision.

The parties may rest assured that this Court carefully considered each and every point of contention proffered and the evidence supportive thereof. The fact that not every party's argument is discussed in detail below should not be taken to mean such argument was not considered.

TENTATIVE DECISION

Sixty years ago, in Brown v. Board of Education (1954) 347 U.S. 483, the United States Supreme Court held that public education facilities separated by race were inherently unequal, and that students subjected to such conditions were denied the equal protection of the laws under the 14th Amendment to the United States Constitution. In coming to its conclusion, the Court significantly noted:

> Today, education is perhaps the most important function of state and local governments. Compulsory school attendance laws and the great expenditures for education both demonstrate our recognition of the importance of education to our democratic society. It is required in the performance of our most basic public responsibilities, even service in the armed forces. It is the very foundation of good citizenship. Today it is a principal instrument in awakening the child to cultural values, in preparing him for later professional training, and in helping him to adjust normally to his environment. In these days, it is doubtful than any child may reasonably be expected to succeed in life if he is denied the opportunity of an education. Such an opportunity, where the state has undertaken to provide it, is a **right** which must be made available to all **on equal terms**.
> Id. at 493 (Emphasis added).

In Serrano v. Priest (1971) 5 Cal.3d 584 (hereinafter Serrano I) and Serrano v. Priest (1976) 18 Cal.3d 728 (hereinafter Serrano II), the California Supreme Court held education to be a "fundamental interest" and found the then-existing school financing system to be a violation of the equal protection clause of the California Constitution, holding that:

> Under the strict standard applied in such (suspect classifications or fundamental interests) cases, the state bears the burden of establishing not only that it has a *compelling* interest which justifies the law but that the distinctions drawn by the law are *necessary* to further its purpose.
> Serrano II, at 761 (quoting Serrano I, at 597 (Original emphasis)).

In Butt v. State of California (1992) 4 Cal.4th 668, the California Supreme Court held that a school district's six-week-premature closing of schools due to revenue shortfall deprived the affected students of their fundamental right to basic equality in public education, noting:

> It therefore appears well settled that the California Constitution makes public education uniquely a fundamental concern of the State and prohibits maintenance and operation of the public school system in a way which denies **basic educational equality** to the students of particular districts. The State itself bears the ultimate authority **and** responsibility to ensure that its district-based system of common schools provides **basic equality of educational opportunity**.
> Id. at 685 (Emphasis added).

What Brown, Serrano I and II, and Butt held was that unconstitutional laws and policies would not be permitted to compromise a student's fundamental right to equality of the educational experience. Proscribed were: 1) Brown: racially based segregation of schools; 2) Serrano I and II: funding disparity; and 3) Butt: school term length disparity. While these cases addressed the issue of a lack of **equality** of education based on the discrete facts raised therein, here this Court is directly faced with issues that compel it to apply these constitutional principles to the **quality** of the educational experience.

Plaintiffs are nine California public school students who, through their respective *guardians ad litem,* challenge five statutes of the California Education Code, claiming said statutes violate the equal protection clause of the California Constitution. The allegedly offending statutes are: 44929.21(b) ("Permanent Employment Statute"); 44934, 44938(b)(1) and (2) and 44944 (collectively "Dismissal Statutes"); and 44955 ("Last-In-First Out (LIFO)"). Collectively, these statutes will be referred to as the "Challenged Statutes".

Plaintiffs claim that the Challenged Statutes result in grossly ineffective teachers obtaining and retaining permanent employment, and that these teachers are disproportionately situated in schools serving predominately low-income and minority students. Plaintiffs' equal protection claims assert that the Challenged Statutes violate their fundamental rights to equality of education by adversely affecting the quality of the education they are afforded by the state.

This Court is asked to directly assess how the Challenged Statutes affect the educational experience. It must decide whether the Challenged Statutes cause the potential and/or unreasonable exposure of grossly ineffective teachers to all California students in general and to minority and/or low income students in particular, in violation of the equal protection clause of the California Constitution.

This Court finds that Plaintiffs have met their burden of proof on all issues presented.

PROCEDURAL HISTORY

This action was filed on May 14, 2012; on August 15, 2012, the currently operative First Amended Complaint for Declaratory and Injunctive Relief was filed against defendants 1)State of California; 2) Edmund G. Brown, Jr., in his official capacity as Governor of California; 3)Tom Torkalson, in his official capacity as State Superintendent of Public Instruction; 4)California Department of Education; 5)State Board of Education (1-5 hereinafter are collectively referred to as "State Defendants"); 6) Los Angeles Unified School District (LAUSD); 7)Oakland Unified School District (OUSD); and 8)Alum Rock Union School District (ARUSD).

On November 9, 2012, this Court, through written opinion, overruled demurrers filed by State Defendants and ARUSD. Thereupon, it indicated that controlling questions of law involving substantial grounds for difference of opinion existed and that appellate resolution may materially advance conclusion of litigation, pursuant to California Code of Civil Procedure 166.1, thus inviting appellate review of its rulings on the demurrers. On

December 10, 2012, Defendants filed a petition for writ of mandate with the Court of Appeal, which issued a stay of all proceedings in this Court on December 18. On January 29, 2013, the Court of Appeal denied the relief requested by Defendants, returning the matter to this Court for further proceedings.

On May 2, 2013, this Court, recognizing the legitimate and immediate interests in this litigation of the California Teachers Association and the California Federation of Teachers (collectively "Intervenors"), granted their respective motions to intervene, thereby allowing them to become fully vested parties herein and allowing the presentation of the **legal positions** of the widest-possible range of interested parties.

(This Court stresses **legal positions** intentionally. It is not unmindful of the current intense political debate over issues of education. However, its **duty and function** as dictated by the Constitution of the United States, the Constitution of the State of California and the Common Law, is to avoid considering the political aspects of the case and focus only on the legal ones. That this Court's decision will and should result in political discourse is beyond question but such consequence cannot and does not detract from its obligation to consider only the evidence and law in making its decision.

It is also not this Court's function to consider the wisdom of the Challenged Statutes. As the Supreme Court of California stated in In re Marriage Cases (2008) 43 Cal.4th 757 at 780:

> It is also important to understand at the outset that our task in this proceeding is not to decide whether we believe, *as a matter of policy*, that the officially recognized relationship of a same-

> sex couple should be designated a marriage rather than a domestic partnership (or some other term), but instead only to determine whether the difference in the official names of the relationships *violates the California Constitution.* (Original emphasis).

While judges of this country and state do not leave their personal opinions at the courthouse door every morning, it is incumbent upon them not to let such opinions color their view of the cases before them that day. The Supreme Court goes on:

> Whatever our views as individuals with regard to this question as a matter of policy, we recognize as judges and as a court our responsibility to limit our consideration of the question to a determination of the constitutional validity of the current legislative provisions. In re Marriage Cases, at 780.)

Plaintiffs voluntarily dismissed with prejudice: 1)ARUSD on September 13, 2013; 2)LAUSD on September 18; and 3)OUSD on December 23.

On December 13, 2013, by written opinion, this Court denied State Defendants'/Intervenors' motions for Summary Judgment/Summary Adjudication. Moving parties sought reversal of this ruling from the Court of Appeal through petition for writ of mandate/prohibition and request for stay of proceedings. This relief was summarily denied by the Court of Appeal on January 14, 2014, thus returning the matter to this Court for further proceedings, including trial.

Trial commenced January 27, 2014. Motions for judgment pursuant to CCP 631.8 made by State Defendants/Intervenors after Plaintiffs rested were denied March 4. The trial concluded with oral argument on March 27 and with final written briefs filed on April 10, at which time the matter stood submitted to this Court for decision.

ANALYSIS

Since the Challenged Statutes are alleged to violate the California Constitution, the pertinent provisions thereof are set forth:

> Article 1, sec. 7(a): "A person may not be deprived of life, liberty, or property without due process of law or denied equal protection of the laws … ."
>
> Article 9, sec. 1: "A general diffusion of knowledge and intelligence being essential to the preservation of the rights and liberties of the people, the Legislature shall encourage by all suitable means the promotion of intellectual, scientific … improvement."
>
> Article 9, sec. 5: "The Legislature shall provide for a system of common schools by which a free school shall be kept up and supported in each district … ."

In Serrano I and II and Butt, *supra*, an overarching theme is paradigmatized: the Constitution of California is the ultimate guarantor of a meaningful, basically equal educational opportunity being afforded to the students of this state.

State Defendants' exhibit 1005, "California Standards for the Teaching Profession" (CSTP)(2009) in its opening sentence declares: "A growing body of research confirms that the **quality of teaching** is what matters most for the students' development and learning in schools." (Emphasis added).

All sides to this litigation agree that competent teachers are a critical, if not the most important, component of **success** of a child's in-school educational experience. All sides also agree that grossly ineffective teachers substantially **undermine** the ability of that child to succeed in school.

Evidence has been elicited in this trial of the specific effect of grossly ineffective teachers on students. The evidence is compelling. Indeed, it shocks the conscience. Based on a massive study, Dr. Chetty testified that a single year in a classroom with a grossly ineffective teacher costs students $1.4 million in lifetime earnings per classroom. Based on a 4 year study, Dr. Kane testified that students in LAUSD who are taught by a teacher in the bottom 5% of competence lose 9.54 months of learning in a single year compared to students with average teachers.

There is also no dispute that there are a significant number of grossly ineffective teachers currently active in California classrooms. Dr. Berliner, an expert called by State Defendants, testified that 1-3% of teachers in California are grossly ineffective. Given that the evidence showed roughly 275,000 active teachers in this state, the extrapolated number of grossly ineffective teachers ranges from 2,750 to 8,250. Considering the effect of grossly ineffective teachers on students, as indicated above, it therefore cannot be gainsaid that the number of grossly ineffective teachers has a direct, real, appreciable, and negative impact on a significant number of California students, now and well into the future for as long as said teachers hold their positions.

Within the framework of the issues presented, this Court must now determine what test is to be applied in its analysis. It finds that based on the criteria set in Serrano I and II and Butt, and on the evidence presented at trial, Plaintiffs have proven, by a preponderance of the evidence, that the Challenged Statutes impose a real and appreciable impact on students' fundamental right to equality of education **and** that they impose a disproportionate burden on poor and minority students. Therefore the

Challenged Statutes will be examined with "strict scrutiny", and State Defendants/Intervenors must "bear[] the burden of establishing not only that [the State] has a *compelling* interest which justifies [the Challenged Statutes] but that the distinctions drawn by the law[s] are *necessary* to further [their] purpose." Serrano I, 5 Cal.3d at 597 (Original emphasis).

PERMANENT EMPLOYMENT STATUTE

The California "two year" statute is a misnomer to begin with. The evidence established that the decision not to reelect must be formally communicated to the teacher on or before March 15 of the second year of the teacher's employment. This deadline already eliminates 2-3 months of the "two year" period. In order to meet the March 15 deadline, reelection recommendations must be placed before the appropriate deciding authority well in advance of March 15, so that in effect, the decision whether or not to reelect must be made even earlier. Bizarrely, the beneficial effects of the induction program for new teachers, which lasts an entire two school years and runs concurrently with the Permanent Employment Statute, cannot be evaluated before the time the reelection decision has to be made. Thus, a teacher reelected in March may not be recommended for credentialing after the close of the induction program in May, leaving the applicable district with a non-credentialed teacher with tenure. State Defendants' PMQ Linda Nichols testified that this would leave the district with a "real problem because now you are not a credentialed teacher; and therefore, you cannot teach." She further opined that State Superintendent of Education Tom Torlakson "clearly believes, you know it would theoretically be great" to have the tenure decision made after induction was over.

There was extensive evidence presented, including some from the defense, that, given this statutorily-mandated time frame, the Permanent Employment Statute does not provide nearly enough time for an informed decision to be made regarding the decision of tenure (critical for both students and teachers). As a result, teachers are being reelected who would not have been had more time been provided for the process. Conversely, startling evidence was presented that in some districts, including LAUSD, the time constraint results in **non**-reelection based on "any doubt," thus depriving 1)teachers of an adequate opportunity to establish their competence, and 2)students of potentially competent teachers. Brigitte Marshall, OUSD's Associate Superintendent for Human Resources, testified that these are "high stakes" decisions that must be "well grounded and well founded."

This Court finds that **both** students and teachers are unfairly, unnecessarily, and for no legally cognizable reason (let alone a **compelling** one), disadvantaged by the current Permanent Employment Statute. Indeed, State Defendants' experts Rothstein and Berliner each agreed that 3-5 years would be a better time frame to make the tenure decision for the mutual benefit of students and teachers.

Evidence was admitted that nation-wide, 32 states have a three year period, and nine states have four or five. California is one of only five outlier states with a period of two years or less. Four states have no tenure system at all.

This Court finds that the burden required to be carried under the strict scrutiny test has not been met by State Defendants/Intervenors, and

thus finds the Permanent Employment statute unconstitutional under the equal protection clause of the Constitution of California. This Court enjoins its enforcement.

DISMISSAL STATUTES

Plaintiffs allege that it is too time consuming and too expensive to go through the dismissal process as required by the Dismissal Statutes to rid school districts of grossly ineffective teachers. The evidence presented was that such time and cost constraints cause districts in many cases to be very reluctant to even commence dismissal procedures.

The evidence this Court heard was that it could take anywhere from two to almost ten years and cost $50,000 to $450,000 or more to bring these cases to conclusion under the Dismissal Statutes, and that given these facts, grossly ineffective teachers are being left in the classroom because school officials do not wish to go through the time and expense to investigate and prosecute these cases. Indeed, defense witness Dr. Johnson testified that dismissals are "extremely rare" in California because administrators believe it to be "impossible" to dismiss a tenured teacher under the current system. Substantial evidence has been submitted to support this conclusion.

This state of affairs is particularly noteworthy in view of the admitted number of grossly ineffective teachers currently in the system across the state (2750-8250), and of the evidence that LAUSD alone had 350 grossly ineffective teachers it wished to dismiss at the time of trial regarding whom the dismissal process had not yet been initiated.

State Defendants/Intervenors raise the entirely legitimate issue of due process. However, given the evidence above stated, the Dismissal Statutes present the issue of *über* due process. Evidence was presented that classified employees, fully endowed with due process rights guaranteed under Skelly v. State Personnel Board (1975) 15 Cal.3d 194, had their discipline cases resolved with much less time and expense than those of teachers. Skelly holds that a position, such as that of a classified or certified employee of a school district, is a property right, and when such employee is threatened with disciplinary action, due process attaches. However, that due process requires a balancing test under Skelly as discussed at pages 212-214 of the opinion. After this analysis, Skelly holds at page 215:

> [D]ue process does mandate that the employee be accorded certain procedural rights before the discipline becomes effective. As a minimum, these preremoval safeguards must include notice of the proposed action, the reasons therefore, a copy of the charges and materials upon which the action is based, and the right to respond, either orally or in writing, to the authority imposing discipline.

Following the hearing of the administrative agency, of course, the employee has the right of a further multi-stage appellate review process by the independent courts of this state to assess whether the factual determinations are supported by substantial evidence.

The question then arises: does a school district classified employee have a lesser property interest in his/her continued employment than a teacher, a certified employee? To ask the question is to answer it. This Court heard no evidence that a classified employee's dismissal process (i.e., a Skelly hearing) violated due process. Why, then, the need for the current tortuous process required by the Dismissal Statutes for teacher dismissals, which has been decried by both plaintiff and defense witnesses? This is

particularly pertinent in light of evidence before the Court that teachers themselves do not want grossly ineffective colleagues in the classroom.

This Court is confident that the independent judiciary of this state is no less dedicated to the protection of reasonable due process rights of teachers than it is of protecting the rights of children to constitutionally mandated equal educational opportunities.

State Defendants/Intervenors did not carry their burden that the procedures dictated by the Dismissal Statutes survive strict scrutiny. There is no question that teachers should be afforded reasonable due process when their dismissals are sought. However, based on the evidence before this Court, it finds the current system required by the Dismissal Statutes to be so complex, time consuming and expensive as to make an effective, efficient yet fair dismissal of a grossly ineffective teacher illusory.

This Court finds that the burden required to be carried under the strict scrutiny test has not been met by State Defendants/Intervenors, and thus finds the Dismissal Statutes unconstitutional under the equal protection clause of the Constitution of California. This Court enjoins their enforcement.

LIFO

This statute contains no exception or waiver based on teacher effectiveness. The last-hired teacher is the statutorily-mandated first-fired one when lay-offs occur. No matter how gifted the junior teacher, and no matter how grossly ineffective the senior teacher, the junior gifted one, who

all parties agree is creating a positive atmosphere for his/her students, is separated from them and a senior grossly ineffective one who all parties agree is harming the students entrusted to her/him is left in place. The result is classroom disruption on two fronts, a lose-lose situation. Contrast this to the junior/efficient teacher remaining and a senior/incompetent teacher being removed, a win-win situation, and the point is clear.

Distilled to its basics, the State Defendants'/Intervenors' position requires them to defend the propostion that the state has a compelling interest in the *de facto* separation of students from competent teachers, and a like interest in the *de facto* retention of incompetent ones. The logic of this position is unfathomable and therefore constitutionally unsupportable.

The difficulty in sustaining Defendants'/Intervenors' position may explain the fact that, as with the Permanent Employment Statute, California's current statutory LIFO scheme is a distinct minority among other states that have addressed this issue. 20 states provide that seniority **may** be considered among other factors; 19 (including District of Columbia) leave the layoff criteria to district discretion; two states provide that seniority cannot be considered, and only 10 states, including California, provide that seniority is the sole factor, or one that must be considered.

This Court finds that the burden required to be carried under the strict scrutiny test has not been met by State Defendants/Intervenors, and thus finds the LIFO statute unconstitutional under the equal protection clause of the Constitution of California. This Court enjoins its enforcement.

EFFECT ON LOW INCOME/ MINORITY STUDENTS

Substantial evidence presented makes it clear to this Court that the Challenged Statutes disproportionately affect poor and/or minority students. As set forth in Exhibit 289, "Evaluating Progress Toward Equitable Distribution of Effective Educators," California Department of Education, July 2007:

> Unfortunately, the most vulnerable students, those attending high-poverty, low-performing schools, are far more likely than their wealthier peers to attend schools having a disproportionate number of underqualified, inexperienced, out-of-field, and ineffective teachers and administrators. Because minority children disproportionately attend such schools, minority students bear the brunt of staffing inequalities.

The evidence was also clear that the churning (aka "Dance of the Lemons) of teachers caused by the lack of effective dismissal statutes and LIFO affect high-poverty and minority students disproportionately. This in turn, greatly affects the stability of the learning process to the detriment of such students.

CONCLUSION

All Challenged Statutes are found unconstitutional for the reasons set forth hereinabove. All injunctions issued are ordered stayed pending appellate review.

In the event a Statement of Decision is requested pursuant to CRC 3.1590(d), Plaintiffs are ordered to prepare a Proposed Statement of Decision and a Proposed Judgment pursuant to 3.1590(f).

Alexander Hamilton wrote in Federalist Paper 78: "For I agree there is no liberty, if the power of judging be not separated from the legislative and executive powers." Under California's separation of powers framework, it is not the function of this Court to dictate or even to advise the legislature as to how to replace the Challenged Statutes. All this Court may do is apply constitutional principles of law to the Challenged Statutes as it has done here, and trust the legislature to fulfill its mandated duty to enact legislation on the issues herein discussed that passes constitutional muster, thus providing each child in this state with a basically equal opportunity to achieve a quality education.

Dated this 10th of June, 2014

Treu, J.

EXHIBIT 18
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness

ALLISON ATTEBERRY,
SUSANNA LOEB AND
JAMES WYCKOFF

WORKING PAPER 90 • FEBRUARY 2013

Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness

Allison Atteberry
University of Virginia

Susanna Loeb
Stanford University

James Wyckoff
University of Virginia

Contents

Acknowledgements ii

Abstract iii

Introduction 1

Background and Prior Literature 2

Data 5

Methods 8

Results 14

Conclusions 24

Figures and Tables 27

Appendices 38

References 43

Acknowledgements

We appreciate helpful comments from Matt Kraft and Eric Taylor on previous versions of the paper. We are grateful to the New York City Department of Education and the New York State Education Department for the data employed in this paper. We appreciate financial support from the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by IES Grant R305A060018 to the American Institutes for Research. The research reported here was also supported by the IES Grant R305B100009 to the University of Virginia. The views expressed in the paper are solely those of the authors and may not reflect those of the funders. Any errors are attributable to the authors.

CALDER working papers have not gone through final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness

Allison Atteberry, Susanna Loeb, and James Wyckoff

CALDER Working Paper No. 90
February 2013

Abstract

There is increasing agreement among researchers and policymakers that teachers vary widely in their ability to improve student achievement, and the difference between effective and ineffective teachers has substantial effects on standardized test outcomes as well as later life outcomes. However, there is not similar agreement about how to improve teacher effectiveness. Several research studies confirm that on average novice teachers show remarkable improvement in effectiveness over the first five years of their careers. In this paper we employ rich data from New York City to explore the variation among teachers in early career returns to experience. Our goal is to better understand the extent to which measures of teacher effectiveness during the first two years reliably predicts future performance. Our findings suggest that early career returns to experience may provide useful insights regarding future performance and offer opportunities to better understand how to improve teacher effectiveness. We present evidence not only about the predictive power of early value-added scores, but also on the limitations and imprecision of those predictions.

Introduction

Teachers vary widely in their ability to improve student achievement, and the difference between effective and ineffective teachers has substantial effects on standardized test outcomes (Rivkin et al., 2005; Rockoff, 2004) as well as later life outcomes (Chetty, Friedman, & Rockoff, 2011). Given the research on the differential impact of teachers and the vast expansion of student achievement testing, policy-makers are increasingly interested in how measures of teacher effectiveness, such as value-added, might be useful for improving the overall quality of the teacher workforce. Some of these efforts focus on identifying high-quality teachers for rewards, to take on more challenging assignments, or as models of expert practice (see for example, teacher effectiveness policies in the District of Columbia Public Schools). Others attempt to identify struggling teachers in need of mentoring or professional development to improve skills (Taylor & Tyler, 2011; Yoon, 2007). Finally, because some teachers may never become effective, some researchers and policymakers are exploring meaningful increases in dismissals of ineffective teaches as a mechanism for improving the overall quality of teachers. One common feature of all of these efforts is the need to establish a system to identify teachers' effectiveness as early as possible in a way that accurately predicts how well these inexperienced teachers might serve students in the long run.

To date, only a little is known about the dynamics of teacher performance in the first five years. The early career period represents a unique opportunity to identify struggling teachers, examine the likelihood of future improvement, and make strategic pre-tenure dismissals to improve teacher quality. In this paper, we explore how teacher value-added measures in the first two years predict future teacher performance. In service of this larger goal, we pursue a set of questions designed to provide policy makers with concrete insight into how well teacher value-added scores from the first two years of a teacher's career would perform as an early signal of how that teacher would develop over the next five years. We use panel data from the New York City Department of Education that follows all new

teachers who began between the 1999-00 and 2006-07 school years to pursue the following research questions:

- To what extent do teachers vary around the mean pattern in returns to experience? We examine the degree of variability in the developmental trajectories of teacher in terms of effectiveness in the early career.
- To what extent do teachers with different initial value-added scores in the first two years exhibit different returns to experience during the first five years, and how well do these initial scores account for variability in future performance?
- To what extent do predictions made based on early value-added scores mischaracterize teachers' future performance?

In what follows we provide some background for the relevance of this research question, as well as a review of existing literature that helps frame our question. We next describe the data from New York City used in the analysis, as well as the analytic approach used to answer these three research questions. We follow with the results organized by question, and conclude.

Background and Prior Literature

Research documents the substantial impact of assignment to a high-quality teacher on student achievement (Aaronson, Barrow, & Sander, 2007; Boyd, Lankford, Loeb, Ronfeldt, & Wyckoff, 2011; Clotfelter et al., 2007; Hanushek, 1971; Hanushek, Kain, O'Brien, & Rivkin, 2005; Harris & Sass, 2011; Murnane & Phillips, 1981; Rockoff, 2004). We also know that teachers are not uniformly effective. The difference between effective and ineffective teachers has substantial effects on short term outcomes like standardized test scores, as well as longer term outcomes such as college attendance, wages, housing quality, family planning, and retirement savings (Chetty et al., 2011).

Despite the variation in teacher effectiveness, teacher workforce policies generally do not acknowledge these disparities in quality, nor do most districts tailor their responses to or compensation for teachers based on performance. In the *Widget Effect,* Weisberg, Sexton, Mulhern, & Keeling, (2009) surveyed twelve large districts across four states and found that no measures of performance were taken into account in recruitment, hiring/ placement, professional development, compensation, granting tenure, retention, or layoffs except in three isolated cases (Weisberg, Sexton, Mulhern, & Keeling, 2009). While evaluation and compensation reform is currently popular, the majority of districts in the U.S. still primarily use teacher educational attainment, additional credentialing, and experience to determine compensation. In addition, while principal observations of teachers is common practice, there is often little useful variation in principals' evaluations of teachers (Weisberg et al., 2009).

Given the growing recognition of the differential impacts of teachers, policy-makers are increasingly interested in how measures of teacher effectiveness such as value-added or observational measures might be useful for improving the overall quality of the teacher workforce. In the field, policy makers rarely propose to use value-added scores as the exclusive metric for teacher evaluation. The Measures of Effective Teaching (MET Project), Ohio's Teacher Evaluation System (TES), and D.C.'s IMPACT policy are all examples where value-added scores are considered in conjunction with other evidence from the classroom, such as observational protocols or principal assessments. In this paper we focus on value-added scores as illustrative of teacher quality measures more broadly, not because we believe that value-added scores should be used in isolation. Practically speaking, there are very few places where other measures of teacher effectiveness are readily available at this point to study a panel of teachers throughout their first five years.

The utility of teacher effectiveness measures for policy use depends on properties of the measures themselves, such as validity and reliability. Measurement work on the reliability of teacher value-added scores has typically characterized reliability using a perspective based on the logic of test-

retest reliability, in which a test administered twice within a short time period is judged based on the equivalence of the results over time. Researchers have thus examined the stability of value-added scores from one year to the next, reasoning that a reliable measure should be consistent with itself from one year to the next (e.g., Aaronson et al., 2007; Goldhaber & Hansen, 2010; Kane & Staiger, 2002; Koedel & Betts, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009). When value-added scores fluctuate dramatically in adjacent years, this presents a policy challenge—the measures may reflect statistical imprecision more than true teacher performance. In this sense, stability is a highly desirable property in a measure of effectiveness, because the conclusions one would draw based on value-added in one year are more likely to be consistent with conclusions made in another year.

It is worth noting that measuring reliability based on stability is potentially more problematic in the first five years of a teacher's career. Inherent in this approach to measuring reliability is the belief that the latent phenomenon of interest—true effectiveness— is not changing over time. Yet, on average, teachers undergo the largest improvements of their careers in the first few years of teaching. Researchers have generally documented a leveling off of returns to experience after five to seven years, suggesting that many teachers reach their own plateau, whatever that may be, during this early career period (Clotfelter, Ladd, & Vigdor, 2006; Clotfelter et al., 2007; Rivkin et al., 2005; Rockoff, 2004).[1] Given that teachers exhibit the largest returns to experience during this early phase, one might expect teacher quality measures to be less stable during this time as evidenced by year-to-year correlation for example. At the same time, these measures may well be reliable in the sense that the scores consistently reflect latent true quality as it develops and, in theory these scores may be just as predictive of future scores despite their instability.

[1] There are clearly higher average student outcomes for students when exposed to teachers with more experience, though there has been more debate about which years are most formative and whether there are no additional returns to experience after a certain point (Papay & Kraft, 2011).

That said, there are many reasons to be skeptical about our ability to make fair and accurate judgments about teachers based on their first one or two years in the classroom. Anecdotally, one often hears that the first two years of teaching are a "blur," and that virtually every teacher is overwhelmed and ineffective. If in fact first-year teachers' effectiveness is more subject to random influences and less a reflection of their true abilities, their early evaluations would be less predictive of future performance than evaluations later in their career. In this paper we explore the how actual value-added scores from new teachers' first two years might be used by policy makers to anticipate the future effectiveness of their teaching force and to identify teachers early in their career for particular human capital responses.

Data

The backbone of the data that we use for this analysis is administrative records from a range of sources including the New York City Department of Education (NYCDOE), the New York State Education Department (NYSED). The combination of sources provides the student achievement data and the link between teachers and students that we need to create measures of teacher effectiveness and growth over time.

New York City students take achievement exams in math and English Language Arts (ELA) in grades three through eight; however, for the current analysis, we restrict the sample to elementary school teachers (grades four and five), because of the relative uniformity of elementary school teaching jobs compared with middle school teaching where teachers specialize. All the exams are aligned to the New York State learning standards and each set of tests is scaled to reflect item difficulty and are equated across grades and over time. Tests are given to all registered students with limited accommodations and exclusions. Thus, for nearly all students the tests provide a consistent assessment of achievement from grade three through grade eight. For most years, the data include scores for 65,000 to 80,000 students in each grade. We normalize all student achievement scores by subject, grade

and year to have a mean of zero and a unit standard deviation. Using these data, we construct a set of records with a student's current exam score and lagged exam score(s). The student data also include measures of gender, ethnicity, language spoken at home, free-lunch status, special-education status, number of absences in the prior year, and number of suspensions in the prior year for each student who was active in any of grades three through eight in a given year. For a rich description of teachers, we match data on teachers from the NYCDOE Human Resources database to data from the NYSED databases. The NYCDOE data include information on teacher race, ethnicity, experience, and school assignment as well as a link to the classroom(s) in which that teach taught each year.

Analytic Sample and Attrition

We explore how measures of teacher effectiveness—value-added scores—change during the early career. To do this, we rely on the student-level data linked to elementary school teachers to estimate teacher value-added. Value-added scores can only be generated for the subset of teachers assigned to tested grades and subjects. In addition, because we herein analyze patterns in value-added scores over the course of the first five years of a teacher's career, we can only include teachers who do not leave teaching before we can observe their later performance. Not only is limiting the sample to teachers with a complete vector of value-added central to the research question, it also addresses a possible attrition problem. The attrition of teachers from our sample threatens the validity of our estimates because we cannot observe how these teachers would have performed had they remained in the profession, and there is some reason to believe that early attriters may have different returns to experience (Boyd, Lankford, Loeb, and Wyckoff, 2007). As a result, our primary analyses focus on the set of New York City elementary teachers who began between 2000 and 2006 who have, at a minimum, value-added scores in all of their first five years.

Despite the advantages to limiting the sample in this way, the restriction introduces a different problem having to do with external validity. If teachers who are less effective leave teaching earlier or are removed from tested subjects or grades, the estimates of mean value-added across the first five years would be biased upward because the sample is limited at the outset to a more effective subset of teachers. That is, teachers who are consistently assigned to tested subjects and grades for five consecutive years may be quite different from those who are not. Given this tradeoff, we conduct sensitivity analyses and present results also for a less restrictive subsample that requires a less complete history of value-added scores.

Table 1 gives a summary of sample sizes by subject and additional requirements based on minimum value-added scores required. There are 7,656 math teachers (7,611 ELA) who are tied to students in NYC, began teaching during the time period in which they could possibly have at least five years of value-added scores, and teach primarily elementary grades during this time. At a very minimum, we must observe teachers with a value-added score in the first year, which in itself limits the math sample to 4,170 teachers (4,180 for ELA). Our primary analytic sample for the paper is the subset of 842 math teachers for whom we observe a value-added score in at least each of her first five years (859 ELA). The sample sizes decrease dramatically as one increases the number of required value-added scores, which demonstrates our limited ability to look much beyond the first five years. The notable decrease in sample size reveals that teachers generally do not receive value-added scores in every school year, and in research presented elsewhere we examine why so few teachers receive value-added over a consecutive panel. Because the requirement of having five consecutive years of value-added scores is somewhat restrictive, we also examine results for the somewhat larger subsample of teachers for whom we can be sure they remain in the New York City teacher workforce for at least the first five years but have value-added scores in their first year and two of the following four years (n=2,068 for math, 2,073 for ELA).

Methods

The overarching analytic approach in this paper is to follow a panel of new teachers as they go through their first five years and retrospectively examine how performance in the first two years predicts performance thereafter. In order to do so, we first estimate yearly value-added scores for all teachers in New York City. We then use these value-added scores to characterize teachers' developing effectiveness over the first five years to answer the research questions outlined above. We begin by describing the methods used to estimated teacher-by-year value-added scores, and then we lay out how these scores are used in the analysis.

Estimation of Value Added

Although there is no consensus about how best to measure teacher quality, in this project we define teacher effectiveness using a value-added framework in which teachers are judged by their ability to stimulate student standardized test score gains. While imperfect, these measures have the benefit of directly measuring student learning and they have been found to be predictive of other measures of teacher effectiveness such as principals' assessments and observational measures of teaching practice (Atteberry, 2011; Grossman et al., 2010; Jacob & Lefgren, 2008; Kane & Staiger, 2012; Kane, Taylor, Tyler, & Wooten, 2011; Milanowski, 2004), as well as long term student outcomes (Chetty et al., 2011). Our methods for estimating teacher value-added are consistent with the prior literature. Equation 1 describes our approach.[2]

[2] To execute the model described in equation (1), we use a modified version of the method proposed by the Value-Added Research Center (VARC). This approach involves a two-stage estimation process, which is intended to allow the researcher to account for classroom characteristics, which are collinear with the teacher-by-experience fixed effects that serve as the value-added models themselves. This group of researchers is currently involved in producing value-added scores for districts such as New York City, Chicago, Atlanta, and Milwaukee (among others). For more information, see http://varc.wceruw.org/methodology.php

$$A_{itgsy} = \beta_0 + A_{itgs,y-1}\beta_1 + A^{other}_{itgs,y-1}\beta_2 + X_{itgsy}\beta_3 + C_{tgsy}\beta_4 + S_{sy}\beta_5 + \pi_g + \theta_{yt} + \varepsilon_{jitgsy} \quad (1)$$

The outcome A_{itgsy} is the achievement of student i, with teacher t, in grade g, in school s, at time y, and we model this as a function of a vector $A_{itgs,y-1}$of that student's prior achievement in the prior year in the same subject and $A^{other}_{itgs,y-1}$in the other subject (math or ELA); the students' characteristics, X_{itgsy}; classroom characteristics, C_{tgsy}, which are the aggregate of student characteristics as well as the average and standard deviation of student prior achievement; $S_{sy}\beta_5$, school time-varying controls, grade fixed effects, π_g; teacher-by-experience fixed effects (θ_{yt}); as well as a random error term, ε_{jitgsy}.[3] The teacher-by-experience fixed effects become the value-added measures which serve as the outcome variable in our later analyses. They capture the average achievement of teacher t's students in year y, conditional on prior skill and student characteristics, relative to the average teacher in the same subject and grade. Finally, we apply an Empirical Bayes shrinkage adjustment to the resulting teacher-by-year fixed effect estimates to adjust for measurement error.

In the model presented above for the estimation of teacher-by-year value-added scores, we have made several important analytic choices about the best specification for our purposes herein. Our preferred model uses a lagged achievement approach wherein a student's score in a given year serves as the outcome, with the prior year score on the right-hand side (as opposed to modeling gain scores as the outcome).[4] We attend to student sorting issues through the inclusion of all available student covariates rather than using student fixed effects, in part because the latter restricts the analysis to

[3] The effects of classroom characteristics are identified from teachers who teach multiple classrooms per year. The value-added models are run on all teachers linked to classrooms from 2000 on, however the analytic sample for this paper is limited to elementary grade teachers.

[4] Some argue that the gain score model is preferred because one does not place any prior achievement scores which are measured with error on the right-hand side, which introduces potential bias. On the other hand, the gain score model has been criticized because there is less variance in a gain score outcome and a general loss of information and heavier reliance on the assumption of interval scaling. In addition, others have pointed out that the gain score model implies that the impacts of interest persist undiminished rather than directly estimating the relationship between prior and current year achievement (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; McCaffrey et al., 2009).

comparisons only between teachers who have taught at least some students in common.[5] At the school level we also opt to control for all observed school-level covariates that might influence the outcome of interest rather than including school fixed effects, since this would also only allow valid comparisons within the same school. In an appendix, we examine results across a variety of value-added models, including models with combinations of gain score outcomes, student, and school fixed effects.

RQ 1. Estimating Mean and Variance in Returns to Experience

We first estimate the mean returns to experience for teachers in the first five years in order to establish that findings from this dataset are consistent with prior literature. Importantly, however, we also consider whether teachers vary around that overall pattern. That is, we look for evidence of variability in the developmental trajectories of teacher in terms of effectiveness in the early career.

Annual student-level test score data provide the base for estimating returns to experience. In creating measures of growth, we tackle common problems researchers face when estimating returns to experience, particularly isolating the impact of experience on student achievement. We estimate teachers' improvement with experience using a standard education production function quite similar to Equation 1 in that both include the same set of lagged test scores, student, classroom, and school covariates, as well as grade fixed effects. We remove teacher-by-experience fixed effects and replace them with experience level and year fixed effects. The coefficients of interest are those on the set of experience variables. If the experience measures are indicator variables for each year of experience, the coefficient on the binary variable that indicates an observation occurred in a teacher's fifth year represents the expected difference in outcomes between students who have a teacher in her first versus

[5] A student fixed effects approach has the advantage of controlling for all observed and unobserved time-invariant student factors, thus perhaps strengthening protections against bias. However, the inclusion of student-level fixed effects entails a dramatic decrease in degrees of freedom, and thus a great deal of precision is lost (see discussion in McCaffrey et al., 2009). In addition, experimental research by Kane and Staiger (2008) suggests that student fixed effects estimates may be *more* biased than similar models using a limited number of student covariates.

fifth year, controlling for all other variables in the model. We plot these estimated coefficients alongside estimates from other research projects since the mean trend has been the focus of considerable prior work.

We are primarily interested in teachers vary around this mean trend. In order to explore this, we randomly sample 100 teachers from our analytic sample and plot their observed value-added scores during their first five years. We also present the standard deviation of estimated value-added scores across teachers at each year of experience to examine whether the variance in teacher effectiveness appears to be widening or narrowing during the early career.

RQ 2. Performance in the Initial Years of Teaching as a Predictor of Future Effectiveness

Our second research question asks: To what extent do teachers with different initial value-added scores in the first two years exhibit different returns to experience during the first five years, and how well do these initial scores account for variability in future performance? To build off our work exploring variability around mean returns to experience, we explore whether one possible source of that variability is differences in teachers' initial effectiveness. We therefore begin by estimating mean value-added score trajectories throughout the first five years separately by quintiles of teachers' initial performance, and we examine the likelihood that teachers transition from low to high quintiles (and vice versa) during the course of their early career. Policy makers often translate raw evaluation scores into four or five performance groups in order to facilitate direct action for top and bottom performers. Because this practice is so ubiquitous, we also adopt this general approach for characterizing early career performance for a given teacher for many of our analyses. (The creation of such quintiles, however, requires non-trivial analytic decisions on our part and thus we delineate some of our challenges in Appendix A. These analytic choices are likely at play for policy-makers in the real world as well, and thus the discussion of our process may be instructive to this larger audience.)

In order to examine how the development of teacher effectiveness during the early career varies by quintile of initial performance, we model the teacher-by-year value-added measures generated by Equation (1) as outcomes using a non-parametric function of experience with interactions for initial quintile. We plot the coefficients on the interactions of experience and quintile dummy variables to illustrate separate mean value-added trajectories by initial quintile.

We are also interested in whether *any* initially high-performing teachers become among the lowest-performing teachers in the future (or vice versa). We therefore also present a quintile transition matrix that tabulates the number of teachers in each initial quintile (rows) by the number of teachers in each quintile of the mean of their following three years (columns), along with row percentages.

It is worth noting that quintile groupings may obscure large differences between teachers at either extreme within the same quintile, or it may exaggerate the differences between teachers just on either side of one of these cut points. For this reason, we present analyses that move away from reliance on quintiles in order to characterize the relationship between continuous measures of initial and future performance among new teachers. We estimate regression models that predict a teacher's continuous value-added score in a future period as a function of a set of her value-added scores in the first two years of teaching. This approach allows us to consider the ability to predict future scores with initial scores without introducing quintile groupings into the analysis.

We use the following equation to predict each teacher's value-added score in a given "future" year (e.g., value-added score in years three, four, five, or the mean of these) as a function of value-added scores observed in the first and second year. We present results across a number of value-added outcomes and sets of early career value-added scores, however Equation (3) describes the fullest specification which includes a cubic polynomial function of all available value-added data in both subjects from teachers' first two years:

$$E\left[VA_{m,y=3,4,5}\right] = \beta_0 + f^3\left(VA_{m,y=1}\right) + f^3\left(VA_{m,y=2}\right) + f^3\left(VA_{e,y=1}\right) + f^3\left(VA_{e,y=2}\right) \quad (3)$$

We summarize results from forty different permutations of Equation (3)—by subject and by various combinations of value-added scores used—by presenting the adjusted R-squared values from each model. This comparison illustrates the proportion of variance in future performance that can be accounted for using early value-added scores, and to easily consider the comparative improvements of using more scores or different scores in combination with one another.

RQ 3. Examining Errors in Prediction

Finally, because we know that errors in prediction are inevitable, we present evidence on the degree of confidence in our predictions, and the nature of the miscategorizations one might make based on value-added scores from a teacher's first two years. We examine confidence intervals around forecasted future scores from the most promising specifications of Equation (4) above. In addition, we present a framework for thinking about the kinds of mistakes likely to be made and for whom those mistakes are costly. We base this framework loosely on the statistical concept of Type I and Type II errors, and we then apply this framework to historical data from New York City. We propose a hypothetical policy mechanism in which value-added scores from the early career are used to rank teachers and identify the strongest or weakest for any given human capital response (be it pay for performance, professional development, probation, dismissal, etc.). We then follow teachers into the following five years and calculate the proportion of the initially identified teachers who actually turn out to be high- or low- effective teachers in the long run.

Results

Mean and Variance in Early Career Improvement by Experience

Researchers consistently have found that, on average, teachers become more effective at improving student test performance during their first few years of teaching. Figure 1 depicts returns to experience from eight studies, as well as our own estimates using data from New York City.[6] Each study shows increases in student achievement as teachers accumulate experience such that by a teacher's fifth year her or his students are performing, on average, from 5 to 15 percent of a standard deviation of student achievement higher than when he or she was a first year teacher. This effect is substantial, given that a one standard deviation increase in teacher effectiveness is typically about 15 percent of standard deviation of student achievement; thus, the average development over the first few years of teaching is from one-third to a full standard deviation in overall teacher effectiveness.[7]

Figure 1 demonstrates that early career teacher experience is associated with large student achievement gains, on average. However, this estimate of average early career improvement may obscure the substantial variation across teachers around this mean trajectory—that is, some teachers may improve a lot over time while others do not. Indeed, we find evidence of substantial variance in value-added to student achievement across teachers. Figure 2 plots the observed value-added score trajectories for 100 teachers who were randomly sampled from the set of New York City elementary teachers that have value-added scores in their first five years (our analytic sample), alongside the mean value-added scores (red) in the same period. This graph illustrates notable variability around the mean

[6] Results are not directly comparable due to differences in grade level, population, and model specification, however Figure 1 is intended to provide some context for estimated returns to experience across studies for our preliminary results.

[7] See Hanushek, Rivkin, Figlio, & Jacob (2010) for a summary of studies that estimate the standard deviation of teacher effectiveness measures in terms of student achievement. The estimates for Reading are between 0.11 and 0.26 standard deviations across studies, while the estimates for math are larger and also exhibit somewhat more variability (0.11 to 0.36, but with the average around 0.18 standard deviations (Aaronson et al., 2007; Hanushek & Rivkin, 2010; Jacob & Lefgren, 2008; Kane, Rockoff, & Staiger, 2008; Thomas J. Kane & D.O. Staiger, 2008; Koedel & Betts, 2011; Nye, Konstantopoulos, & Hedges, 2004; Rivkin et al., 2005; Rockoff, 2004; Rothstein, 2010).

growth during this time period, which suggests that the mean returns to experience may not characterize individual teachers well.

To further explore variation in returns to experience, we calculate the standard deviation of teacher value-added scores across teachers within each year of experience for both the complete analytic sample and the teachers randomly selected for Figure 2. For English Language Arts (ELA) the standard deviations in teacher value-added is 0.20 across teachers in their first year (experience = 0). For math, the standard deviation of first-year teacher value-added is approximately 0.21. The variance in both ELA and math value-added scores steadily increase with experience so that the standard deviation in value added is at least 0.23 by the fifth year of teaching, representing an increase of 15 to 30 percent from the first year. The trends suggest that the processes associated with teacher development create greater differences in teaching effectiveness over these early years of teaching and, thus, that there is likely to be meaningful variation in returns to experience across teachers.

Performance in the Initial Years of Teaching as a Predictor of Future Effectiveness

One way to make sense of the substantial variability observed above is to examine mean value-added scores over years of experience separately by quintiles of initial performance. If initial performance provides insight into future performance, we should see that the highest quintile of initial performance continue to be the highest performing quintile over time (and vice versa for the initially lowest quintile). We group teachers by initial performance quintiles of the mean of their first *two* years. Figure 3 plots mean value-added scores by experience for each quintile of performance in the first two years among teachers with value-added scores in at least the first five years. (See Appendix for a series of checks using different samples of teachers based on minimum years of value-added scores required, definitions of initial performance quintiles, and specifications of the value-added model.)

Figure 3 provides evidence of consistent differences in value-added across quintiles of initial performance. On average, the initially lowest-performing teachers are consistently the lowest-performing, the highest are consistently the highest. While the lowest quintile does exhibit the most improvement, this set of teachers does not, on average, "catch up" with other quintiles, nor are they typically as strong as the median first year teacher even after five years. However, the mean trajectories by quintile shown in Figure 3 may obscure further important within-quintile variance. That is, it provides little information about movements across quintiles in the future. In Table 3, we present a quintile transition matrix that tabulates the number of teachers in each initial quintile (rows) by the number of teachers in each quintile of the mean of their following three years (columns), along with row percentages.[8] The majority—62 percent—of the initially lowest quintile math teachers ultimately show up in the bottom two quintiles of future performance. On the other end, the initially highest-performing teachers exhibit even more consistency: About 73 percent of these teachers remain in the top two quintiles of mean math performance in the following years. Movements from one extreme to the other are comparatively rare. About 19 percent of bottom- and 10 percent of top- quintile initial performers end up in the opposite extreme two quintiles. Results are similar for ELA teaching.

Taken together, the transition matrix in Table 3 and the results in Figures 1-3 begin to provide a picture of how teachers improve over the first five years. First, consistent with prior findings this is a period of growth overall. Second, in the face of this overall trend, we also observe considerable variability in the patterns of development during this time frame, as evidenced by the plots of individual teachers in Figure 2 and the depiction of quintile-based trajectories in Figure 3. Finally, despite this variability, the transition matrix suggests that measures of value-added in the first two years predict of future performance for most teachers. We next pin down more carefully the extent to which initial performance can provide accurate and meaningful predictions about teacher performance in the future.

[8] We use the mean of years 3, 4, and 5 rather than just the fifth year to absorb some of the inherently noisy nature of value-added scores over time.

In Table 4, we present adjusted R-squared values from a various specifications of Equation (4) above, and we present results across five possible sets of early career value-added scores to explore the additional returns to using more value-added scores. One evident pattern is that additional years of value-added predictors improve the predictions of future value-added—particularly the difference between having one score and having two scores. The lowest adjusted R-squared values come from models that predict a value-added score in one future year using one value-added score from a single prior year. For example, teachers' math value-added scores in the first year only explains 8.9 percent of the variance in value-added scores in the third year. The predictive power is even lower for ELA (2.9 percent). A second evident pattern in Table 4 is that value-added scores from the second year are typically two- to three times stronger predictors than value-added in the first year for both math and ELA.

Recall that elementary school teachers are unique in that they typically teach both math and ELA every year and thus we can estimate both a math and an ELA score for each teacher in each year. When we combine all available value-added scores from both subjects in both of the first two years, and also include cubic polynomial terms for theses scores, we can explain more variance in future scores. Table 4 also shows that the measure of future score is as important as the measure of initial score. Initial scores do a far better job of predicting a teachers' average value-added over a group of years than of predicting value-added in any of the individual years. For math, when including all first and second year value-added measures, we explain about 27.8 percent of the variance in average future performance compared with no more than 19.4 percent of the variance in any of the individual future years. (For ELA, the comparable results are 20.9 percent and 15.4 percent.)

Table 4 shows early scores can explain up to a fifth of the variation in future scores; however, it is not necessarily clear whether this magnitude is relatively big or relatively small. For comparison, we estimate the predictive ability of measured characteristics of teachers during their early years. These

include typically available measures: indicators of a teacher's pathway into teaching, available credentialing scores and SAT scores, competiveness of undergraduate institution, teacher's race/ ethnicity, and gender. When we predict math mean value-added scores in years three through five using this set of explanatory factors, we explain only 2.8 percent of the variation in the math outcome (2.5 percent for ELA).[9] The measured teacher characteristics that district leaders typically have at their disposal to predict who will be the most or least effective teachers clearly do not perform as well as value-added scores from the first two years.

Potential Errors in Categorizing Teachers

The prior analyses provide evidence that initial performance is predictive of later performance; however, the analyses also imply that this predictive ability is far from perfect. In this section we further describe the error associated with these predictions. To provide one perspective on our ability to predict future value-added scores, we return to Equation (4) above, in which we model mean value-added scores in years three through five as cubic polynomial functions of value-added scores in both subjects in the first two years. Using this model, we can predict future performance and present a conservative confidence interval for each forecasted prediction point (see Figure 4).

As Figure 4 shows, even 80 percent confidence intervals are quite large for individual predictions. The mean squared error for teachers in this sample is about 0.14, which is approximately equivalent to a standard deviation in the overall distribution of teacher effectiveness. The degree of error for individual predictions is substantively large, and we can see that teachers' predicted future value-added scores differ markedly from the observed scores based on distance from the y=x line. That said, recall that the adjusted r-squared from this simple model of future performance is high—about 27.8 percent of the variance in future performance can be accounted for using value-added scores in the

[9] These results not shown, available upon request.

first and second years. Certainly the value-added based predictions of future performance are imprecise, and accordingly most policy makers argue that value-added scores should not be used in isolation to reward or sanction teachers. Nonetheless, the movement towards a more strategic approach to human capital management in the K-12 setting drives us to consider the utility of the tools at hand in light of the current lack of strong alternatives on which to base predictions of how teachers will serve students throughout their career.

A policy that uses value-added scores to group teachers based on performance will likely produce groups that are not entirely distinct from one another in future years. Figure 5 presents the complete distribution of future value-added scores by initial quintile. These depictions provide a more complete sense of how groups based on initial effectiveness overlap in the future.[10] For each group, we have added two reference points. First, the "+" sign located on each distribution represents the mean of future performance in each respective initial-quintile group. The color-coded vertical lines represent the mean *first* year performance by quintile. This allows the reader to compare distributions both to where the group started on average, as well as to where other groups have ended up on average in future years.

The vast majority of policy proposals based on value-added target teachers at the top (for rewards, mentoring roles, etc.) or at the bottom (for support, professional development, or dismissal). Thus, even though the middle quintiles are not particularly distinct in Figure 4, it is most relevant that the top and bottom initial quintiles are. In both math and ELA, there is some overlap of the extreme quintiles in the middle—some of the initially lowest-performing teachers appear to be just as skilled in future years as initially high-performing teachers. However, the majority of these two distributions are distinct from one another.

[10] The value-added scores depicted in each distribution are each teacher's mean value-added score in years three, four, and five. For brevity, we refer to these scores as "future" performance.

We can take a closer look at the initially lowest quintile of performance relative to some meaningful comparison points. For example in math, the large majority (76.5 percent) of the density of the lowest (red) quintile lies to the left of the mean of the distribution of future scores for the middle quintile (the comparable percentage is 74.4 percent for ELA). Thus, most of the initially lowest performers never match the performance of an average fifth year teacher (of course this implies that about a quarter of the initially-lowest performing quintile—those who appear at the very top of the red distribution of future performance— do surpass the mean of the middle quintile).

Figure 5 also allows us to compare the distribution of initially lowest quintile math teachers to the average teacher in the first year of experience (yellow vertical line), as this is the expected performance of a teacher with whom one could replace a dismissed teacher. It turns out that 68.9 percent of math teachers do not exceed the comparison to the average first year teacher (66.6 percent for ELA). In addition, an ineffective teacher retained for three additional years imposes three years of below-average performance on students. The longer a teacher with low true impacts on students is retained, the expected differential impact on students will be the *sum* of the difference between an average new teacher and the less effective teacher across years of additional retention.

This discussion lends itself naturally to a consideration of the tradeoffs associated with identifying teachers as low-performing based on imperfect measurements from a short period of time in the early career. The goal is to maximize the percentage of teachers for whom we accurately predict future performance based on early performance. There are two possible errors—Type I and Type II—that one could make in service of this goal. We begin with the null hypothesis that a given teacher is *not* ineffective in the long run (for the sake of clarity, think of this as assuming a teacher is at least average). In this case, a Type I error is rejecting a true null hypothesis, which is to falsely identify a teacher as low-performing when she turns out to be at least average in the long run. This type of error typically dominates the value-added debate, because this error negatively and unfairly penalizes teachers who

would be dismissed even though they *would have* emerged as effective over time. On the other hand, Type II error is often overlooked even though it likely affects students' instructional experiences. In the case of Type II error, one fails to reject a false null hypothesis, which implies that one fails to identify a teacher as ineffective when she actually is ineffective in the long run. This error might be quantified as the percentage of teachers who perform poorly in the future who were not identified as low-performing based on initial performance. Depending on the definition of ineffective, students who are assigned to teachers who persist as a result of Type II error receive a lower quality of instruction than they would have had the teacher been replaced by an average new teacher.

While we have framed the discussion of Type I and Type II error in terms of identifying ineffective teachers, a parallel approach can be taken to identifying excellent teachers. In this case, the null hypothesis is that a given teacher is *not* excellent in the long run. Type I error is rejecting a true null hypothesis—thinking that a teacher will be excellent when he or she is not. Type II error is not rejecting the null when it is true—thinking that a teacher will not be excellent when he or she is. To the extent that excellent teachers deserve recognition, Type II error in this context impacts teachers. To the extent that by identifying excellent teachers schools can improve their quality of instruction, Type I error, in this context, impacts students.

In practice, identifying Type I and Type II errors is complex, in part because it requires a clear criterion for identifying future "ineffectiveness" or "excellence". The measures we have of future quality are imprecise; narrow, as they are based only on student test performance in math and ELA; and relative instead of absolute, as they compare teacher to each other rather than to a set standard. We have ameliorated to some extent the measurement error in a teacher's value-added measure in a given year by (1) using Bayes shrunk estimates which attenuates extreme measures in proportion to their imprecision, (2) averaging across multiple future years to lessen the influence of any one outlier result, and (3) breaking effectiveness into quintiles, so that while teachers in the middle quintiles may be less

distinct, one can focus on teachers at the extremes of future performance using top and bottom quintiles. We, however, do not address the narrowness of the value-added measure, nor its relative nature.

Figure 6 helps to illustrate Type I and Type II error associated with identifying teachers as ineffective, perhaps for the purpose of dismissal. In practice, this same approach could be used for any number of strategic policy responses such as allocating additional support, mentoring, observation, or professional development. Simply for the clarity of the example, we describe a dismissal policy. We start by translating the mean value-added scores of teachers in years one and two into percentiles. Moving from left to right along the x-axis represents an increase in the percentage of teachers who are identified as ineffective, and as a result might be dismissed. The y-axis gives the corresponding percent from each of the top and bottom three *future* deciles (separate lines for each decile) that would be dismissed based on the x-axis value. For example, a vertical line at 10 would give the percent of each of the future deciles that would be identified as ineffective if we were too identify the 10 percent of the lowest value-added teachers in the first two years as ineffective.

We can garner a great deal of information from this figure. First, it is clear that while there are errors in identifying ineffective teachers even when initial ineffectiveness is defined at a very low level (e.g. the 5th percentile), most of the teachers identified end up in the bottom part of the distribution of future performance. Second, not surprisingly, the errors get bigger as we aim to identify a higher proportion of teachers as ineffective. For example, a substantial portion of teachers in the bottom 50 percent of initial value added end up in the top three deciles of future value-added.

To make the example more concrete, consider a hypothetical dismissal policy of the bottom ten percent of teachers in initial value-added. In this case, we are attempting to test a hypothesis about whether a teacher will be ineffective or not (the null hypothesis). We see that this policy would eliminate 29.5 percent of teachers who would subsequently appear in the lowest decile of future

performance and another 22.1 percent of teachers who would appear in the second lowest decile. In contrast, none of the top decile of future performance would be (falsely) identified and only two percent of the second highest decile would be (falsely) identified. The latter two numbers can also be thought of as a quantification of the Type I error—teachers who were identified as low performing by the policy but ultimately appeared to be among the highest performers in the future.

Figure 6 also illustrates Type II error. At the ten percent threshold, while 29.5 percent of the lowest decile teachers would have been dismissed, the other 70.5 percent of the lowest decile were not (fail to reject a false null). If one believes that the bottom ten percent of the distribution of performance in years three through five is a good criteria for ineffectiveness, then the failure to identify these teachers can be viewed through the lens of Type II error.[11] As one moves to the right on the x-axis, dismissing a larger proportion of teachers based on initial value-added, these tradeoffs balance one another. At 20 percent dismissal rate, one loses half (51.5 percent) of the future bottom decile in math (and fails to eliminate the other half of that quintile), while the relative "cost" is 6.8 percent of the top decile.[12]

One could argue about the appropriate criteria for future effectiveness. Another reasonable assertion might be to characterize every teacher who is significantly less effective than an average teacher and then retained as a Type I error, and every teacher who becomes significantly more effective than an average teacher who is accidentally dismissed as a Type II error (a more extreme interpretation of these results). We are agnostic about what should be used by policy makers in practice as the "right" criteria.

---

[11] Of course, the ability to eliminate a large percentage of the bottom deciles of future performance is capped by the percentage of teachers one is willing to fire. Put more concretely, if one adopts a policy of firing the bottom five percent of teachers after the first two years, even a "perfect" measure could only dismiss at most 50 percent of the bottom decile (i.e., 5 percent of the whole sample equals 50 percent of one decile).

[12] It is worth noting that, at some point, firing an unreasonably high percentage of teachers may trigger a general equilibrium problem, and the assumption that there is a continuous supply of "average" new teachers will no longer be true. The further to the right we move along the x-axis in Figure 6, not only is the likelihood of making type I errors much greater, but the likelihood of encountering a shortage of qualified teachers also increases.

Conclusions

From a policy perspective, the ability to predict future performance is practically most important for inexperienced teachers because policies that focus on development (e.g. mentoring programs), dismissal, and promotion are likely most relevant during this period. Prior work has documented the relationship between a teacher's value-added in one year and his or her value-added in future years. These analyses have been based on teachers without any restriction based on teaching experience. However, there is reason to believe that the relationship between current performance and future performance might be different for novice teachers than for other teachers. In particular, substantial evidence suggests that on average teachers improve more (that is, change their performance more) over the first years of teaching than over subsequent years.

In this paper we describe the trajectory of teachers' performance over their first five years as measured by their value-added to ELA and math test scores of students and how this trajectory varies across teachers. Our goal is to assess the potential for predicting future performance (performance in years 3, 4, and 5) based on teachers' performance in their first two years. We focus particularly on Type I and Type II error where Type I error is falsely classifying teachers into a group to which they do not belong (e.g. ineffective or excellent) and Type II error is failing to classify teachers into a group to which they belong.

We find that, on average, initial performance is quite predictive of future performance, far more so than measured teacher characteristics such as their own test performance (e.g. SAT) or their educational experience. On average the highest fifth of teachers remain the highest fifth of teachers; the second fifth remains the second fifth; the third fifth remains the third fifth; and so on. Predictions are particularly powerful at the extremes. Initially excellent teachers are far more likely to be excellent teachers in the future than are teachers who were not as effective in their first few years.

This said, any predictions we make about teachers' future performance are far from perfect. The predicted future scores we estimated were, on average, about 0.14 standard deviation units off from actual scores (RMSE), which represents a substantial range of possible effectiveness. Certainly, when it comes to making policy based on any imprecise measures of teacher effectiveness, there is no avoiding that some mistakes will be made. Thinking about these errors using the lens of Type I versus Type II errors emphasizes the fact that there are tradeoffs to be made in practice. While most attention has been paid to the former—falsely identifying teachers as ineffective when they ultimately are not—the latter represents the failure to identify and address teaching that does not serve students well in terms of their academic outcomes. The paper highlights the balance between these two kinds of error and also sheds light on how complex it is to definitively know when these mistakes are made.

We see three immediate strands coming out of the work completed herein. First, we will expand our existing analysis to middle school teachers. There are reasons to believe that the training, structure, and organization of middle schools might produce a different growth experience than observed in the elementary teacher population. Indeed some preliminary work suggests that the relationship between initial and future performance may be less straight forward in higher grades.

The second next step that arises from this work is to examine potential causes for the notable variability in growth rates in the early career. While the most effective teachers tend to remain the most effective and the least effective remain among the least effective, Figure 2 depicts a wide range of developmental patterns across the teachers in the first five years. Moreover, even when we break the mean value-added score trajectories over time into quintiles, there is undoubtedly important within-quintile variation. That is, even among the initially least effective teachers, some make up more ground than others. In future work we seek to identify correlates of teachers' growth over this time period. Our interest in this work is piqued by a variance decomposition of the growth in teacher effectiveness over the first five years of teaching indicating that 30 percent of the variance lies between schools, and 70

percent within schools. In our larger dataset, we observe a great deal about how teachers were trained, measures of their generic and teaching abilities, educational attainment details, and pathways into teaching. Further, once teachers begin teaching, they are undoubtedly influenced by (for better or worse) the organizational nature of the schools to which they are assigned, their colleagues, school leaders, and opportunities for professional development. For a few cohorts of New York City teachers, we can also look more deeply at the experiences of new teachers using in-depth survey data for teachers who have recently completed their first year. Work in this area is intended to help district leaders and policy makers understand how new teacher experiences might be modified to improve the quality of the existing teacher workforce.

Finally, we are interested in an observation that arose as an artifact of trying to follow teachers across multiple years with value-added scores: Of the 5,516 elementary math teachers who began teaching in or after the 1999-2000 school year and were present in the teacher database for at least their first five years, only 842 (about 15.3 percent) received value-added scores in every year. Some preliminary work suggests to us that teachers who possessed more value-added scores during their early career tended to be somewhat higher-performing in their initial year. Certainly there are a number of reasons that could account for missing value-added scores—e.g., switching to a non-tested subject or grade, insufficient numbers of tested students in a given year, leaves of absence. It is also possible that some of those explanations could be systematic or strategic on the part of teachers and principals. While that behavior is in itself of interest to those who wish to understand how teachers and schools might respond to evaluation policies, it is also interesting to note that we can evaluate such policies only for teachers who have at least some minimum amount of consistent evidence about how they perform over time. To the extent that this represents a somehow selective sample, the conclusions we reach about these policies may be less generalizable to all teachers. We think that examining the nature of the data patterns that arise in a district like New York might be instructive to the larger field.

Figures

Figure 1:
Student Achievement Returns to Teacher Early Career Experience, Preliminary Results from Current Study (Bold) and Various Other Studies

Results are not directly comparable due to differences in grade level, population, and model specification, however Figure 1 is intended to provide some context for estimated returns to experience across studies for our preliminary results. Current= Results for grade 4 & 5teachers who began in 2000+ with at least 9 years of experience. For more on model, see Technical Appendix. C,L V 2007= = Clotfelter, Ladd, Vigdor (2007; Rivkin, Hanushek, & Kain, 2005), Table 1, Col. 1 & 3; P, K, 2011 = Papay & Kraft (2011), Figure 4 Two-Stage Model; H, S 2007 = Harris & Sass (2011), Table 3 Col 1, 4 (Table 2); R, H, K, 2005= Rivkin, Hanushek, Kain (2005), Table 7, Col. 4; R(A-D) 2004 = Rockoff (2004), Figure 1 & 2, (A= Vocab, B= Reading Comprehension, C= Math Computation, D= Math Concepts); O 2009 = Ost (2009), Figures 4 & 5 General Experience; B,L,L,R,W 2008 = Boyd, Lankford, Loeb, Rockoff, Wyckoff (2008).

Figure 2:
Variance across Teachers in Quality (VA) over Experience, by Subject and Attrition Group.

Supplement to Figure 2.
Standard Deviation of Estimated Value Added Scores, by Levels of Experience in Figure 2
(Across All Teachers in the Sample, versus 100 Teachers Randomly Sampled for the Figure)

	Math					*ELA*				
	E= 0	E=1	E=2	E=3	E=4	E= 0	E=1	E=2	E=3	E=4
Full Sample	0.215	0.231	0.236	0.242	0.240	0.204	0.214	0.222	0.228	0.229
100 Teachers	0.211	0.232	0.230	0.243	0.241	0.192	0.204	0.220	0.231	0.230

Figure 3:
Mean VA Scores, by Subject (Math or ELA), Quintile of Initial Performance, and Years of Experience for Elementary School Teachers with VA Scores in at Least First Five Years of Teaching.

Figure 4:
Predicted Future Value-Added Scores (Mean of Years, 3,4, and 5) based on Observed Valued-Added Scores in Years 1 and 2, by Actual Future Value-Added Scores, with 80% Confidence Intervals Around Individual Predictions.

Figure 5:
Distribution of Future Value-Added Scores, by Initial Quintile of Performance

Figure 6:
Departures by Future Performance Quintile Based on Early Career Performance

Tables

Table 1:
Analytic Sample Sizes by Cumulative Restrictions

	Math		ELA	
	# Tchrs	**# Obs**	**# Tchrs**	**# Obs**
All Grade 4-8 Teachers Tied to Students in NYC since 2000	16,909	45,979	17,607	47,753
Started Teaching in 2000- 2006	13,355	39,367	13,942	41,041
Modal Grade in First Five Years is Grade 4 or 5	7,656	24,219	7,611	24,282
In HR Dataset for At Least 5 Years	5,516	20,790	5,482	20,860
Has VA Score in At Least 1st Year	4,170	14,085	4,180	14,226
Has VA in 1st and at Least 2 of Next 4 Years	2,068	10,853	2,073	10,967
Has VA in At Least Years 1 thru 3	1,792	9,544	1,798	9,642
Has VA in At Least Years 1 thru 5	842	5,685	859	5,822
Has VA in At Least Years 1 thru 7	329	2,780	346	2,918
Has VA in At Least Years 1 thru 9	135	1,324	139	1,362

Table 2:
Difference in Mean Value Added and Numbers of Final Analytic Sample Teachers in each Quintile of Initial Performance, by Approach to Quintile Construction

		Q1	Q2	Q3	Q4	Q5
Math Quintiles….						
… of All Teacher-Years (1)	n	104	158	207	219	154
	mean	-0.114	0.010	0.099	0.180	0.310
... After Limiting to Teachers in First Year (2)	n	51	125	165	242	259
	mean	-0.204	-0.069	0.031	0.132	0.306
… And Limiting to Elementary Teachers (3)	n	145	158	185	174	180
	mean	-0.125	0.011	0.102	0.188	0.346
… And Limiting to Teachers with 5+ VA score (4)	n	169	168	169	168	168
	mean	-0.112	0.028	0.113	0.196	0.354
ELA Quintiles…						
… of All Teacher-Years (1)	n	137	171	185	235	131
	mean	-0.107	-0.022	0.066	0.139	0.253
... After Limiting to Teachers in First Year (2)	n	81	127	179	236	236
	mean	-0.201	-0.079	0.008	0.100	0.258
… And Limiting to Elementary Teachers (3)	n	168	168	192	172	159
	mean	-0.144	-0.015	0.068	0.151	0.298
… And Limiting to Teachers with 5+ VA score (4)	n	172	172	172	172	171
	mean	-0.142	-0.012	0.067	0.145	0.291

Note: We construct quintiles of performance in a teacher's first two years. The final analytic sample of teachers is restricted to the teachers who taught primarily fourth or fifth grade and for whom we observe at least five consecutive years of VA scores, beginning in the teacher's first year of teaching. Note that method (3) above is the preferred approach for this paper.

Table 3. Quintile Transition Matrix from Initial Performance to Future Performance, by Subject (Number, Row Percentage, Col Percentage)

		Quintile of Future Performance					
Math Initial Quintile		Q1	Q2	Q3	Q4	Q5	Row
Q1	n	53	37	28	17	10	145
	(row %)	(36.6)	(25.5)	(19.3)	(11.7)	(6.9)	
	(col %)	(47.3)	(23.4)	(14.1)	(8.5)	(5.7)	
Q2	n	23	37	49	38	11	158
	(row %)	(14.6)	(23.4)	(31.0)	(24.1)	(7.0)	
	(col %)	(20.5)	(23.4)	(24.7)	(19.0)	(6.3)	
Q3	n	22	45	50	42	26	185
	(row %)	(11.9)	(24.3)	(27.0)	(22.7)	(14.1)	
	(col %)	(19.6)	(28.5)	(25.3)	(21.0)	(14.9)	
Q4	n	10	25	40	55	44	174
	(row %)	(5.7)	(14.4)	(23.0)	(31.6)	(25.3)	
	(col %)	(8.9)	(15.8)	(20.2)	(27.5)	(25.3)	
Q5	n	4	14	31	48	83	180
	(row %)	(2.2)	(7.8)	(17.2)	(26.7)	(46.1)	
	(col %)	(3.6)	(8.9)	(15.7)	(24.0)	(47.7)	
	Column Total	112	158	198	200	174	**842**
		Quintile of Future ELA Performance					
ELA Initial Quintile		Q1	Q2	Q3	Q4	Q5	Row
Q1	n	49	45	40	23	11	168
	(row %)	(29.2)	(26.8)	(23.8)	(13.7)	(6.5)	
	(col %)	(40.8)	(23.7)	(20.0)	(11.7)	(7.2)	
Q2	n	33	54	39	28	14	168
	(row %)	(19.6)	(32.1)	(23.2)	(16.7)	(8.3)	
	(col %)	(27.5)	(28.4)	(19.5)	(14.2)	(9.2)	
Q3	n	19	43	48	57	25	192
	(row %)	(9.9)	(22.4)	(25.0)	(29.7)	(13.0)	
	(col %)	(15.8)	(22.6)	(24.0)	(28.9)	(16.4)	
Q4	n	9	37	45	45	36	172
	(row %)	(5.2)	(21.5)	(26.2)	(26.2)	(20.9)	
	(col %)	(7.5)	(19.5)	(22.5)	(22.8)	(23.7)	
Q5	n	10	11	28	44	66	159
	(row %)	(6.3)	(6.9)	(17.6)	(27.7)	(41.5)	
	(col %)	(8.3)	(5.8)	(14.0)	(22.3)	(43.4)	
	Column Total	120	190	200	197	152	**859**

Note: Initial quintiles are constructed by first restricting the sample to grade four and five teachers and then identifying five equally sized groups based on a teacher's mean value-added score in her first two years. The quintiles of future performance are constructed by first restricting the sample to grade four and five teachers and then identifying five equally-sized groups based on a teacher's mean value-added score in years three, four, and five. The sample is subsequently restricted to teachers with value-added scores in at least the first five years.

Table 4:
Adjusted R-Squared Values for Regressions Predicting Future (Years 3, 4, and 5) VA Scores as a Function of Sets of Value-Added Scores from the First Two Years

	Outcome			
Early Career VA Predictor(s)	VA in Y3	VA in Y4	VA in Y5	Mean(VA_{Y3-5})
Math				
Math VA in Y1 Only	0.089	0.052	0.070	0.109
Math VA in Y2 Only	0.153	0.165	0.141	0.241
Math VA in Y1 & Y2	0.178	0.171	0.158	0.265
VA in Both Subjects in Y1 & Y2	0.179	0.188	0.166	0.277
VA in Both Subjects in Y1 & Y2 (cubic)	0.175	0.194	0.172	0.278
ELA				
ELA VA in Y1 Only	0.029	0.049	0.023	0.064
ELA VA in Y2 Only	0.062	0.114	0.069	0.154
ELA VA in Y1 & Y2	0.075	0.135	0.077	0.181
VA in Both Subjects in Y1 & Y2	0.090	0.145	0.087	0.203
VA in Both Subjects in Y1 & Y2 (cubic)	0.094	0.154	0.086	0.209

Table 5. Year-to-Year Correlations, by Subject and Experience

	Y+0	Y+1	Y+2	Y+3	Y+4	Y+5
Math						
All Teachers	1.000	0.436	0.386	0.343	0.308	0.291
New Teachers	1.000	0.373	0.328	0.288	0.246	0.175
ELA						
All Teachers	1.000	0.327	0.291	0.247	0.223	0.239
New Teachers	1.000	0.230	0.181	0.168	0.145	0.167

Note: The table presents pairwise correlations between value-added scores in a given year (Y) and the subsequent year (Y+1), two years later (Y+2), … , five years later (Y+5). We do this for two samples of teachers—the full sample of teachers who teach elementary grades in New York City (without regard to years of experience), and a subsample of teachers who began their career in year Y.

Appendix A

The most straightforward approach to making quintiles would be to simply break the full distribution of teacher-by-year fixed effects into five groups of equal size. However, we know that value-added scores for first year teachers are, on average, lower than value-added scores for teachers with more experience. For the purposes of illustration, imagine that first year teacher effects comprise the entire bottom quintile of the full distribution. In this case, we would observe no variability in first year performance—that is, all teachers would be characterized as "bottom quintile" teachers, thus eliminating any variability in initial performance that could be used to predict future performance. We thus chose to center a teacher's first year value-added score around the mean value-added for first year teachers and then created quintiles of these centered scores. By doing so, quintiles captured whether a given teacher was relatively more or less effective than the average *first* year teacher, rather than the average teacher in the district.

In order to trace the development of teachers' effectiveness over their early career, we limited the analytic sample to teachers with a complete set of value-added scores in the first five years. As is evident from Table 1 above, relatively few teachers meet this restrictive inclusion criterion. We hesitated to first restrict the sample and then make quintiles solely within this small subset, because we observed that teachers with a more complete value-added history tended to have higher initial effectiveness. In other words, a "bottom quintile" first year teacher in the distribution of teachers with at least five consecutive years of value-added might not be comparable to the "bottom quintile" among all first years teachers for whom we might wish to make predictions. For this reason, we made quintiles relative to the sample of all teachers regardless of the number of value-added scores they possessed, and subsequently limited the sample to those with at least five years of value-added. As a result of this choice, we observe slightly more top quintile teachers than bottom quintile teachers in the initial year. However by making quintiles before limiting the sample, we preserve the absolute thresholds for those

quintiles and thus ensure that they are consistent with the complete distribution of new teachers. In addition, it is simply not feasible for any districts to make quintiles in the first year or two depending on how many value-added scores *will* have in the first five years.

Finally, our ultimate goal is to use value-added information from the early career to produce the most accurate predictions of future performance possible. Given the imprecision of any one year of value-added scores, we average a teacher's value-added scores in years one and two and make quintiles thereof. We present some specification checks by examining our main results using value-added from the first two years in a variety of ways (e.g., first year only, second year only, a weighted average of the first two years, teachers who were consistently in the same quintile in both years). In Table 2, we present the number of teachers and mean of value-added scores in each of five quintiles of initial performance, based on these various methods for constructing quintiles. One can see that the distribution of the teachers in the analytic sample (fourth and fifth grade teachers with value-added scores in first five years) depends on quintile construction.

Appendix B

In Figure 3 of the paper, we present mean value-added scores over the first five years of experience, by initial performance quintile. Here we recreate these results across three dimensions: (A) minimum value-added required for inclusion in the sample, (B) how we defined initial quintiles, and (3) specification of the value-added models used to estimate teacher effects:

(A) We examine results across two teacher samples based on minimum value-added required for inclusion. The first figure uses the analytic sample used throughout the main paper—teachers with value-added scores in at least all of their first five years. The second widens the analytic sample to the set of teachers who are consistently present in the dataset for at least five years, but only possess value-added scores in their 1st, and 2 of the next 4 years.

(B) We examine results across four possible ways of defining quintiles: (1) "Quintile of First Year"—this is quintiles of teachers' value-added scores in their first year alone; (2) "Quintile of the Mean of the First Two Years"—this is quintiles of teacher's *mean* value-added scores in the first *two* years and is the approach we use throughout the paper; (3) "Quintile Consistent in First Two Years"—here we group teachers who were consistently in the same quintiles in first and second year (i.e., top quintile both years); and (4) "Quintile of the Mean of Y1, Y2, & Y2"—the quintiles of teacher's mean value added score in first and second year, double-weighting the second year.

(C) Finally, we examine results using two alternative value-added models to the one used in the paper. "VA Model B" uses a gain score approach rather than the lagged achievement approach used in the paper. "VA Model D" differs from the main value-added model described in the paper in that it uses student-fixed effects in place of time-invariant student covariates such as race/ ethnicity, gender, etc. See next page for results.

Elem Teachers with VAM in All of First Five Years

Quintile Among All 1st-Yr Tchrs

Quintile Of Mean of First 2 Years

Quintile Consistent in First Two Years

Quintile Of Mean of Y1, Y2, Y2

Elem Teachers with VAM in 1st, and 2 of Next 4 Years

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1).

Atteberry, A. (2011). *Stacking up: Comparing Teacher Value Added and Expert Assessment*. Working Paper.

Boyd, D. J., Lankford, H., Loeb, S., Rockoff, J. E., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management, 27*(4), 793-818.

Boyd, D. J., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). The role of teacher quality in retention and hiring: Using applications to transfer to uncover preferences of teachers and schools. *Journal of Policy Analysis and Management, 30*(1), 88-110.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. National Bureau of Economic Research.

Clotfelter, C., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778.

Clotfelter, C., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Goldhaber, D., & Hansen, M. (2010). Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. Center for Education Data and Research.

Grossman, P. L., Loeb, S., Cohen, J., Hammerness, K. M., Wyckoff, J., Boyd, D. J., & Lankford, H. (2010). Measure for Measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores. *NBER Working Paper*.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review, 61*(2), 280-288.

Hanushek, E. A., Kain, J., O'Brien, D., & Rivkin, S. (2005). The market for teacher quality. *NBER Working Paper*.

Hanushek, E. A., & Rivkin, S. G. (2010). Constrained Job Matching: Does Teacher Job Search Harm Disadvantaged Urban Schools? : National Bureau of Economic Research.

Hanushek, E. A., Rivkin, S. G., Figlio, D., & Jacob, B. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7), 798-812.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615-631.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*, 91-114.

Kane, T. J., & Staiger, D. O. (2008). *Are Teacher-Level Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates*. Working Paper. Retrieved from http://isites.harvard.edu/fs/docs/icb.topic245006.files/Kane_Staiger_3-17-08.pdf

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching, Measures of Effective Teaching Project: Bill and Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources, 46*(3), 587-613.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. University of Missouri Department of Economics Working Paper, (708).

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18-42.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67.

McCaffrey, D. F., Sass, T. R., Lockwood, J., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53.

Murnane, R., & Phillips, B. (1981). What do effective teachers of inner-city children have in common?* 1. *Social Science Research, 10*(1), 83-100.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Ost, B. (2009). How Do Teachers Improve? The Relative Importance of Specific and General Human Capital.

Papay, J. P., & Kraft, M. A. (2011). Do Teachers Continue to Improve with Experience? Evidence of Long-Term Career Growth in the Teacher Labor Market. *Working Paper*.

Rivkin, S., Hanushek, E. A., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement*. *Quarterly Journal of Economics, 125*(1), 175-214.

Taylor, E. S., & Tyler, J. H. (2011). The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers: National Bureau of Economic Research.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect. *Brooklyn, NY: The New Teacher Project*.

Yoon, K. S. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, US Dept. of Education.

EXHIBIT 19
TO
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

TEACHER LAYOFFS: AN EMPIRICAL ILLUSTRATION OF SENIORITY VERSUS MEASURES OF EFFECTIVENESS

Donald Boyd
Center for Policy Research
University at Albany
Rockefeller College
Albany, NY 12222
boydd@rockinst.org

Hamilton Lankford
School of Education
University at Albany
Albany, NY 12222
hamp@albany.edu

Susanna Loeb
School of Education
Stanford University
Stanford, CA 94305
sloeb@stanford.edu

James Wyckoff
(corresponding author)
Curry School of Education
University of Virginia
Charlottesville, VA 22903
wyckoff@virginia.edu

Abstract

School districts are confronting difficult choices in the aftermath of the financial crisis. Today, the financial imbalance in many school districts is so large that there may be few alternatives to teacher layoffs. In nearly all school districts, layoffs are currently determined by some version of teacher seniority. Yet, alternative approaches to personnel reductions may substantially reduce the harm to students from staff reductions relative to layoffs based on seniority. As a result, many school district leaders and other policy makers are raising important questions about whether other criteria, such as measures of teacher effectiveness, should inform layoffs. This policy brief, a quick look at some aspects of the debate, illustrates the differences in New York City public schools that would result if layoffs were determined by seniority in comparison to a measure of teacher effectiveness.

© 2011 Association for Education Finance and Policy

INTRODUCTION

School districts are confronting difficult choices in the aftermath of the financial crisis. In prior recessions, districts often muddled through by imposing a combination of tax increases and expenditure cuts that avoided involuntary personnel reductions. Today the financial imbalance in many school districts is so large that there is no alternative to teacher layoffs. In nearly all school districts, layoffs are currently determined by some version of teacher seniority. Yet alternative approaches to personnel reductions may substantially reduce the harm to students from staff reductions relative to layoffs based on seniority. First, because salaries of novice teachers are often much lower than those of veteran teachers, seniority-based layoffs lead to more teachers being laid off to meet any given budget deficit, with the associated implications for class size. Second, because teachers vary substantially in their effectiveness, staff reduction polices that do not consider effectiveness likely will allow some ineffective teachers to continue teaching while some more effective teachers lose their jobs. Third, because many districts have redesigned human resource policies to place greater emphasis on the recruitment and retention of effective teachers, they may have hired disproportionately more effective teachers over the last several years than in prior years. In such cases, seniority-based layoffs will be even more detrimental for quality. Finally, if, as in many districts, novice teachers are concentrated in schools serving low-achieving students and students in poverty, a seniority-based layoff approach will disproportionately affect the students in those schools. As a result, many school district leaders and other policy makers are raising important questions about whether other criteria, such as measures of teacher effectiveness, should inform layoffs.[1]

This policy brief, a quick look at some aspects of the debate, illustrates the differences in New York City public schools that would result if layoffs were determined by seniority in comparison to a measure of teacher effectiveness. Due to data limitations and an interest in simplicity, our analysis employs the value added of teachers using the fourth- and fifth-grade math and English language arts (ELA) achievement of their students. Unsurprisingly, we find that layoffs determined by a measure of teacher effectiveness result in a more effective workforce than would be the case with seniority-based layoffs. However, we were surprised by facets of the empirical results. First, assuming readily available measures of teacher effectiveness actually measure true teacher effectiveness, an assumption to which we return below, the differences between

1. For a summary of reactions by policy makers to seniority-based layoffs, see Sawchuck 2010. The National Council on Teacher Quality (2010) and the New Teacher Project (2010) raise concerns about relying solely on seniority when making layoffs. Sepe and Roza (2010) find that inexperienced teachers are much more likely to be found in schools with concentrations of poor and low-performing students, thus concentrating the effect of seniority-based layoffs in those schools.

seniority- and effectiveness-based layoffs are larger and more persistent than we anticipated. Second, even though seniority-based layoffs imply laying off more teachers, the differential effect on class size is very small in our simulations, though it would be larger for larger budget reductions. Third, there is somewhat greater school-level concentration of layoffs in a seniority-based system, though with a few notable exceptions both methods result in fairly dispersed layoffs, with the vast majority of schools having no more than one layoff in grades 4 and 5 combined.

So where does this leave us? As a result of the limited applicability of teacher value-added measures (VAMs) to the full population of teachers as well as concerns about potential mismeasurement of effectiveness associated with using VAMs even when available, neither seniority nor measures of value added to student achievement should be the sole criterion determining layoffs. However, ignoring effectiveness measures completely, as seniority-based systems do, is also problematic. Instead, the use of multiple measures of effectiveness for layoff decisions holds promise for softening the detrimental effect of layoffs.

SENIORITY VERSUS EFFECTIVENESS: THE CONCEPTUAL ISSUES

There is substantial evidence that, on average, teachers become more effective over the first few years of their careers (Rivkin, Hanushek, and Kain 2005; Boyd, Lankford et al. 2008; Goldhaber and Hansen 2010). Since in many urban school districts newly hired teachers represent roughly 5 percent of the workforce, a seniority-based layoff that targets less than 10 percent of teachers would eliminate teachers with two or fewer years of experience, typically those teachers who are least effective on average.

However, while teachers typically improve over their first two years, there are some very effective new teachers and some quite ineffective teachers with far greater experience.[2] Seniority-based layoffs result in promising, inexperienced teachers losing their positions while their ineffective but more senior peers continue to teach. As a result, seniority-based layoffs meant to meet budget shortfalls are more detrimental to students than would be a system that laid off the least effective teachers first. However, schools and districts are not able to judge teacher effectiveness perfectly, so the important question is whether they can judge it well enough to improve upon seniority-based layoffs.

2. For example, using North Carolina data, Goldhaber and Hansen (2010) estimate that on average teachers in their fifth year of teaching have math effect sizes that are about 7 percent larger than they had as first-year teachers. However, with a standard deviation of 0.11, there is substantial overlap between the effectiveness of first- and fifth-year teachers.

Several approaches exist to measure teacher effectiveness, including statistical estimates of teacher value added to student achievement, validated observation protocols that are administered by principals or trained evaluators, and less formal observational procedures. Little is known about how these measures overlap or complement each other, but recent work suggests that structured observational protocols correlate with VAMs (Kane et al. 2009; Grossman et al. 2010). In addition, studies find that when asked to evaluate the effectiveness of individual teachers to improve student achievement, principals typically identify as their least effective teachers many of the same teachers identified as least effective by value-added analysis (Jacob and Lefgren 2008; Harris and Sass 2009). These recent studies provide some evidence of overlap among the different approaches to measuring teacher effectiveness.

The value-added approach to measuring teacher effectiveness is gaining popularity, though in most districts principal ratings of teachers are still the only formal measure of teacher effectiveness (see Kane et al. 2009 for a description of the Cincinnati school district, which is an exception to this pattern). Value-added measures have the obvious appeal of linking teachers to the improvement of their students on state-sanctioned exams. In addition, VAMs are generally not as costly to collect as observational measures, especially if a system of standardized testing is already in place. In addition, observational protocols require agreement on what good teaching looks like—agreement that is often difficult to come by. There are, however, a number of well-documented problems with employing value added to evaluate individual teachers in high-stakes decisions such as layoffs. The following are among the more troublesome issues:[3]

- VAMs can be estimated for just those teachers in tested grades and subjects, typically math and reading in grades 4–8.
- While tested outcomes are important, most would agree that effective teachers do more than just improve outcomes as measured on standardized achievement tests.
- Value-added estimates are unstable when based on relatively small numbers of students, thus requiring several classes of students to reduce measurement error.
- Empirically isolating the effect of individual teachers from other school inputs and other attributes, many of which are difficult to measure, is

3. For a more detailed description of issues associated with value added, see Hanushek and Rivkin 2010.

very tricky in the context of comparing teacher effectiveness across diverse school environments, as is the case in many school districts.

Measures of value added either compare teachers within the same school and thus are not able to tell whether on average teachers in one school are more effective than teachers in another school, or they compare teachers across schools and thus may be attributing to the teacher some of the differences in student achievement gains due to schoolwide effects. A potential solution to this issue of separating the school effect from the teacher effect is the use of validated observational protocols, such as CLASS or the Danielson rubric. These protocols measure teacher practices directly and thus are less likely to attribute influences outside the classroom to the teacher. However, the properties of principal or observational protocols are not well developed in the context of high-stakes outcomes such as layoffs.

Choosing the criteria for layoffs is not easy. Using either seniority or currently available measures of effectiveness as the primary determinants suffers from a variety of potential conceptual issues. Seniority suffers from the obvious pitfalls of basing retention on a variable that is only loosely connected to student outcomes. On the other hand, current research is very thin on the properties of any of the effectiveness measures for application in the high-stakes, politically charged layoff environment. Moreover, VAMs suffer from the practical and conceptual limitations described above, while the other measures of effectiveness have either not been validated or require considerable investment to implement.

Nonetheless, many school districts will need to lay off teachers this year. In the remainder of this policy brief we employ data from New York City to simulate the differential effects of layoffs determined by seniority and by VAMs of teacher effectiveness as a means of providing empirical guidance to the following questions:

- Who would be laid off under each approach?
- How does the effect of layoffs on student achievement compare across approaches?
- How does each approach affect schools and class size?
- Are some students disproportionately affected by either approach?

NEW YORK CITY LAYOFFS: AN EMPIRICAL ILLUSTRATION

As an illustration of the differences between layoffs determined by seniority and those determined by value added, we examine the implications of needing to meet a budget shortfall equivalent to 5 percent of total teacher salaries. We

chose 5 percent because this is consistent with discussions in many school districts regarding the magnitude of potential layoffs.[4] In New York City layoffs are determined by inverse seniority in a teacher's license area—for example, childhood education or secondary math.[5] The law provides no guidance on how layoffs should be determined across license areas. Because of the current limitations in the availability of value added and the within-license area requirement of the seniority rules, we apply the budget shortfall to fourth- and fifth-grade teachers, nearly all of whom have a license in childhood education and for whom value added in math and ELA can be estimated. We further structure our analysis by assuming that layoffs applied to teachers as of the summer of 2009 because the data to calculate teacher value added for the 2009–10 teaching workforce are not yet available.

Our primary analysis employs measures of teacher value added and seniority provided by the New York City Department of Education. Unless otherwise noted, all estimates are for teachers who taught during the 2008–9 school year. A teacher's value-added effectiveness estimate is based on up to four years of data, depending on the teacher's longevity within New York City public schools teaching in fourth and fifth grades, with student achievement scores in math and ELA.[6]

We also separately estimate several different models of teacher value added to explore the robustness of the results across different model specifications. The results of these analyses are nearly the same and are noted more specifically below. Our methods for estimating teacher value added are consistent with what is typically found in the literature. We estimate models that include student, classroom, and school variables in addition to teacher fixed effects. We also estimate models with and without controls for teacher experience. Either approach can be appropriate depending on the policy goal. Excluding experience controls ignores the well-documented finding that on average teachers improve over the first four or five years of their career, and less effective novice teachers in 2009 will likely become more effective teachers within just a few years. Controlling for experience adjusts for this difference in effectiveness on average but does not identify the teachers currently most effective. We examine outcomes both ways to assess how much of a difference adjusting for experience makes in practice. The estimates presented below have also been adjusted for the instability associated with measurement error—for example,

4. For example, New York City was recently considering laying off 4,400 teachers, which is about 5.6 percent of teachers (Medina 2010). However, earlier in the year some districts were projecting layoffs of 10 percent or more of their teachers (Lewin and Dillon 2010).
5. Article 52, section 2588 of New York State Education Law defines how layoffs are to be made in New York City.
6. For teachers with fewer than four years of available data, estimates are based on available data.

when a teacher's value-added estimate is based on relatively few students—by employing an empirical Bayes shrinkage adjustment. The appendix outlines our empirical approach in more detail.

The first goal of the analyses is to determine which teachers would be laid off. In order to do this, we assume the budget shortfall that needs to be addressed by layoffs in the fourth and fifth grades is equivalent to a 5 percent reduction in total salaries paid to fourth- and fifth-grade teachers. We then identify which teachers would be laid off to meet this salary reduction under a seniority-based system and which teachers would be laid off to meet this salary reduction under a system that based layoffs on teacher effectiveness as measured by each of the VAMs available—New York City's (NYC) own measure and the ones that we construct. Students take exams in both math and ELA, so we have measures of teachers' value added in both of these subject areas. For the analyses, we average each teacher's value added based on his or her students' math and ELA achievements. We describe the number of teachers laid off overall, the number of teachers from each school laid off, and the average effectiveness of the teachers laid off under each approach.

Only part of the teacher effectiveness that value-added approaches measure is persistent from one year to the next. To address this imperfect persistence, we reestimate value added for the 2006–7 school year, model layoffs as if they happened after that year, and then estimate value added only using the 2007–8 and the 2008–9 data to see how effective the laid-off teachers would have been in the two following years under each approach for determining layoffs. We present the results of these analyses below.

Who Is Laid Off?

Because seniority-based layoffs target teachers whose salaries are typically among the lowest, more layoffs are required to meet any given budget deficit. We find that:

- Layoffs that produce a 5 percent reduction in salaries for NYC fourth- and fifth-grade teachers in 2009 imply terminating about 7 percent of teachers when seniority is the criterion and about 5 percent of teachers when their value-added effectiveness determines terminations. Said slightly differently, in our simulation a layoff system based on value-added results in about 25 percent fewer layoffs than one based on seniority.
- Few of the teachers identified to be laid off are the same under the two approaches. In our simulation, approximately 13 percent of the teachers who are identified to be laid off under a seniority-based system would also be laid off if value added were the criterion. When we employ value-added estimates that control for experience, the comparable statistic is 5 percent.

How Are Schools Affected?

Most schools are simulated to lose relatively few teachers, although some schools would lose more teachers than others. Under a seniority layoff system, 73 percent of schools lose fewer than 10 percent of their fourth- and fifth-grade teachers. However, 12 percent of schools lose more than 20 percent of teachers under the seniority system. Put a little differently, 35 of 708 schools lose three or more fourth- and fifth-grade teachers. Value-added layoffs are somewhat less concentrated. Seventy-two percent of schools have no layoffs, while fewer than 8 percent lose more than 20 percent of their teachers. In this case only 10 of 708 schools lose three or more teachers. In some schools, layoffs would cause important staffing issues, regardless of which system was employed. In these instances, reallocations across schools would likely occur.

What Is the Effect on Class Size?

Reducing the teaching workforce implies an increase in class size, other things equal. However, because our simulation reduces the teaching workforce by only between 5 and 7 percent, class sizes on average increase by fewer than two students, and the average difference between the two methods of laying off teachers is about half a student per class. Nonetheless, as described above, a small portion of schools would lose a meaningful portion of their teachers, and without reallocations across schools this would result in noticeable increases in class size. For example, schools that lost 20 percent of their fourth- and fifth-grade teachers would experience average class size increases of about 5.5 students per class in those grades. It is very likely that teachers would be reallocated across schools to moderate these effects.

How Do Layoffs Affect Achievement?

Figure 1 shows the average math and ELA value-added distribution of all fourth- and fifth-grade teachers and that for teachers laid off under each system. While a system of seniority-based layoffs does terminate some low-value-added teachers (the portion of the left tail of the distribution to the left of the vertical dashed line labeled δ), most of the teachers who would be laid off by seniority have substantially higher value added than even the highest value-added teacher terminated under the value-added criterion. The distribution of teachers laid off under a seniority-based system is very similar to the overall distribution of teacher value added. To the extent that VAMs reflect actual effectiveness in the classroom, the value-added approach identifies the least effective teachers. The typical teacher who is laid off under a value-added system is 26 percent of a standard deviation in student achievement less effective than the typical

Figure 1. Frequencies of Teacher Layoffs by Teacher Value-Added to Math and English Language Arts Achievement

seniority-based layoff.[7] This is a large effect, corresponding to more than twice the difference between a first- and fifth-year teacher and equivalent to the difference between having a teacher who is 1.3 standard deviations below the effectiveness of the average teacher.

We estimate that teachers laid off under the seniority system are much less experienced and have somewhat lower value added than those who remain. Under such a system, those laid off are on average seven years less experienced and have approximately 5 percent of a standard deviation in student achievement lower value added than their peers who remain. As expected, the differences under a layoff system determined by value added are more striking. Using a value-added system, those laid off and their peers who remain differ in experience by about half a year. However, the average value added of those laid off is 31 percent of a standard deviation of student achievement lower than that for teachers who remain.

We can also compare the resulting layoffs to an observational measure of teacher effectiveness. Each year NYC principals rate a small percentage of

7. These differences are very similar to those we obtain when experience controls are included in the estimates of teacher fixed effects.

teachers as "unsatisfactory."[8] Of the teachers in our simulation, 2.5 percent received a U rating between 2006 and 2009. Principal unsatisfactory ratings are much more closely aligned with value-added layoffs than seniority-based layoffs, although not nearly perfectly so. Of teachers who received a U over the last four years, 16 percent would have been laid off under the value-added criterion, while none would have been laid off using the seniority criterion. Teachers in our simulation identified as unsatisfactory by their principals had an average value added of 9 percent of a standard deviation of student achievement lower than their peers who did not receive an unsatisfactory rating. Value-added estimates of effectiveness and the principal ratings have some overlap but address different dimensions of performance.

Because the proportion of all teachers who are terminated under either system is relatively small, depending on one's perspective, the difference between the two systems may or may not be viewed as consequential. For the students who would have been taught by the teachers laid off under each system, the layoff rule is clearly of great consequence. However, because only a relatively small percentage of teachers are laid off, the difference does not have a large effect on the average achievement of students in the district. Under the two scenarios, the average value-added effectiveness of the fourth- and fifth-grade workforce is estimated to differ by less than 2 percent of a standard deviation of student achievement, or about one-tenth of a standard deviation in teacher value added. The size of this effect is about equivalent to 12 percent of the difference in value added between the average first-year teacher and the average fifth-year teacher. While small on average, this effect has some overall impact, and, if lower value-added teachers also reduce the effectiveness of other teachers in their school, the difference between the seniority-based and the value-added systems may be bigger.

The Effect on Future Achievement

Employing value added or other measures of effectiveness for determining layoffs assumes that these measures are good predictors of future effectiveness. However, there are reasons why this might not be the case. For example, there is research documenting that student achievement test scores reflect substantial measurement error (Boyd, Grossman, et al. 2008) and that estimates of teacher value added vary substantially over time (McCaffrey et al. 2009; Goldhaber and Hansen 2010). In addition, despite controlling for experience, some novice teachers will improve more quickly or more slowly than the average. Each of these would suggest that value-added estimates employed

8. The details of the system are set out in "Teaching for the 21st Century" (see www.uft.org/files/attachments/teaching-for-the-21st-century.pdf).

in layoff decisions made one year may mischaracterize teachers' value added in future years.

To explore the effects of layoffs on student achievement in subsequent years, we simulate how layoffs would have been made in the summer of 2007 had the district used data for 2005–6 and 2006–7.[9] We then follow these teachers for two additional years and compare their effectiveness estimates based on the first two years with separate estimates based solely on the second two years in order to assess whether there are persistent differences in effectiveness between outcomes when layoffs are based on seniority rather than value added. This simulation indicates that in 2007 the teachers laid off under a value-added system are, on average, less effective than those laid off under a seniority-based system by 36 percent of a standard deviation of student achievement, which is about 1.9 standard deviations of teacher value added, results that are somewhat greater than our estimates for the 2009 layoff. We follow both groups over the next two years and assess their effectiveness in 2009 using data for just 2008 and 2009. The difference is now 12 percent of a standard deviation of student achievement—equivalent to the difference between first- and fifth-year teachers—and is also equal to having a teacher who is about 0.7 standard deviation less effective than the average teacher. Although there is an important decline in the difference between seniority and value-added estimates when we project the effect to future student achievement, meaningful differences remain. The method by which teachers are laid off has important implications for future achievement.

The comparisons of future achievement presented above represent conservative estimates of the difference between seniority and value-added-based layoffs. When we employ estimates of teacher value added that control for experience, the difference in future achievement between seniority and value-added-based layoffs is 19 percent of a standard deviation in student achievement (compared with the 12 percent shown above), which is 1.0 standard deviation in teacher value added (compared with 0.7) and now equivalent to about 1.5 times the difference between a first- and a fifth-year teacher. In addition, if instead of using combined math and ELA value-added estimates our layoff decisions had been made based solely on math achievement, the difference in value added between seniority-based and value-added-based estimates is 21 percent of a standard deviation in student achievement, which is 1.0 standard deviation of teacher value added. Finally, if the pre- and post-layoff estimates

9. Because we do not have a seniority measure in 2007, we simulate teacher layoffs in that year by laying off all first-year teachers (8.4 percent of all teachers). This resulted in a salary savings of 6.2 percent. We then laid off the least effective teachers until we realized the same salary savings. In this case 6.8 percent of teachers would be laid off, with the most effective laid-off teacher having a value added of −0.28.

were each based on more than two years of data, the reduction in estimation error likely would lead to a higher correlation between those estimates. In turn, the difference between the impact of seniority-based and value-added-based layoffs on future achievement would likely be larger.

Are Some Students Disproportionately Affected?

Teachers often sort systematically into schools based on student characteristics. For example, schools with a higher proportion of black students or students in poverty may have teachers with weaker credentials or less experience (Jackson 2009; Lankford, Loeb, and Wyckoff 2002). As a result, the different criteria for layoffs may differentially affect groups of students. However, our simulation shows little difference in the average characteristics of students taught by teachers laid off under a seniority-based system compared with a value-added approach. For example, teachers laid off based on seniority came from schools where about 80 percent of the students were free or reduced price lunch eligible. The comparable figure for teachers identified for layoff under a value-added criterion taught in schools where 79 percent of the students, on average, were free or reduced price eligible. Similarly, the teachers who would be laid off under a seniority-based rule taught in schools where, on average, 4 percent of the students achieve at level 1 on the fourth-grade math exam, while 4.5 percent of the students performed at level 1 for teachers identified to be laid off by a value-added criterion.

SUMMARY

In the face of substantially diminished revenues, policy makers must juggle a variety of issues in deciding how to close budget deficits. In this regard, policy makers in many school districts believe that teacher layoffs are an important option but struggle with choosing the best criteria for laying off teachers. The standard approach in most school districts relies on measures of seniority. Our simulations show substantial differences in the teachers laid off under a seniority-based system and those who would be laid off if the system instead relied on teacher VAMs. Results for other districts, or even for other grades or license areas in New York City, may differ from those presented here, so this analysis needs to be expanded.

Value added is currently feasible only for the portion of teachers who teach in tested grades and subjects, often math and ELA in grades 4–8, thus limiting its applicability. In addition, as described above, we know little about the extent to which VAMs employing standardized achievement tests capture other important dimensions of teaching. While these issues should be considered in the application of value added, we should not lose sight of the main point. Informing teacher layoffs with measures of effectiveness, while not perfect,

does offer the potential to meaningfully improve the quality of instruction in some classrooms.

Evidence is emerging that several forms of teacher evaluation identify many of the same teachers as least effective. Principal evaluations identifying the least effective teachers overlap with those identified by value-added estimates (Jacob and Lefgren 2008; Harris and Sass 2009). In addition, rigorous teacher observation protocols are positively correlated with VAMs of effectiveness (Grossman et al. 2010; Kane et al. 2009). Given these findings and the large differences found in our layoff simulation, employing fair and rigorous measures of teacher effectiveness for teacher layoffs, rather than seniority measures, can be expected to yield much stronger achievement outcomes for students. Measures that include a variety of approaches of assessing teacher effectiveness offer promise but should be carefully evaluated to better understand their strengths, weaknesses, and complementarities.

We are grateful to the New York City Department of Education and the New York State Education Department for the data employed in this policy brief. We are grateful for financial support from the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by IES Grant R305A060018 to the Urban Institute. The views expressed in the article are solely those of the authors and may not reflect those of the funders. Any errors are attributable to the authors.

REFERENCES

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2008. Measuring effect sizes, the effect of measurement error. CALDER Working Paper No. 19.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah Rockoff, and James Wyckoff. 2008. The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management* 27(4): 793–818.

Carlin, Bradley P., and Thomas A. Louis. 1996. *Bayes and empirical Bayes methods for data analysis*. London: Chapman and Hall.

Goldhaber, Dan, and Michael Hansen. 2010. Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. CALDER Working Paper No. 31.

Grossman, Pamela, Susanna Loeb, Julia Cohen, Karen Hammerness, James Wyckoff, Donald Boyd, and Hamilton Lankford. 2010. Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. NBER Working Paper No. 16015.

Hanushek, Eric A., and Steven G. Rivkin. 2010. Using value-added measures of teacher quality. CALDER Brief No. 9, May.

Harris, Douglas N., and Tim R. Sass. 2009. What makes for a good teacher and who can tell? CALDER Working Paper No. 30.

Jackson, C. Kirabo. 2009. Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics* 27(2): 213–56.

Jacob, Brian A., and Lars Lefgren. 2008. Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics* 26(1): 101–36.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2009. Identifying effective classroom practices using student achievement data. NBER Working Paper No. 15803.

Lankford, Hamilton, Susanna Loeb, and James Wyckoff. 2002. Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis* 24(1): 37–62.

Lewin, Tamar, and Sam Dillon. 2010. Districts warn of deeper teacher cuts. *New York Times*, 20 April.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. The intertemporal variability of teacher effect estimates. *Education Finance and Policy* 4: 572–606.

Medina, Jennifer. 2010. New N.Y. schools face extra pain from layoffs. *New York Times*, 1 June.

National Council on Teacher Quality. 2010. *Teacher layoffs: Rethinking "last-hired, first-fired" policies*. February. Washington, DC: National Council on Teacher Quality.

New Teacher Project. 2010. A smarter teacher layoff system: How quality-based layoffs can help schools keep great teachers in tough economic times. Brooklyn, NY: New Teacher Project.

Rivkin, Steven G., Eric A. Hanushek, and John Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73(2): 417–58.

Sawchuck, Stephen. 2010. Congress urged to tie aid in jobs bill to elimination of seniority-based firing. *Education Week*, 19 May.

Sepe, Cristina, and Marguerite Roza. 2010. The disproportionate impact of seniority-based layoffs on poor, minority students. Seattle, WA: Center on Reinventing Public Education.

APPENDIX: TEACHER VALUE-ADDED ESTIMATION

DATA

The data we employ on teachers in grades 4 and 5 and their students in grades 3, 4, and 5 for school years 2005–6 through 2008–9 come from the New York City Department of Education (NYCDOE). The student data consist of a demographic data file and an exam data file for each year from 2004–5

through 2008–9. The demographic files include measures of gender, ethnicity, language spoken at home, free lunch status, special education status, number of absences, and number of suspensions for each student who was active in grades 3–8 that year. The exam files include, among other things, the year in which an exam was given, the grade level of the exam, and each student's scaled score on the exam. Using these data, we construct a student-level database where exam scores are normalized for each subject, grade, and year to have a zero mean and unit standard deviation, to accommodate any year-to-year or grade-to-grade anomalies in the exam scores. For our purpose, we consider a student to have value-added information in cases in which he or she has a score in a given subject (ELA or math) for a particular year and a score for the same subject in the immediately preceding year for the immediately preceding grade. We do not include cases in which a student took a test for the same grade two years in a row or where a student skipped a grade.

Data on teachers come from the NYCDOE. We employ NYCDOE information to match teachers to their classrooms and students. For this analysis, we employ information on teachers' seniority and experience provided by the NYCDOE and teacher salaries from the New York State Department of Education. Classroom data for students come largely from aggregating student-level data to the classroom level. In addition, we include class size and employ school-level information regarding enrollment, student sociodemographics, and pupil-teacher ratios drawn from the Common Core of Data.

ESTIMATION

Our primary analysis—the simulation of layoffs in 2009—employs measures of teacher value added provided to us by the NYCDOE. The 2007 simulation employs our own value-added estimates. Using student data for the 2005–6 and 2006–7 school years, the effects of student, classroom, and school variables on student achievement are estimated after sweeping out teacher fixed effects. In turn, the teacher fixed effects were estimated by calculating the mean student-level residuals (by teacher) from the first-stage regression. The standard errors for the estimated teacher effects are proxied using the standard deviations of the mean residuals, an approach that ignores the fact that those residuals are based on estimated parameters. We follow the same procedure in creating teacher fixed effect estimates using data for 2006–9, with the obvious difference that these estimates employ four years of data.

Having to use estimates of the actual value teachers add to student achievement, rather than their true value added, implies that it is important to take

into account the corresponding measurement error. We follow the standard approach of adjusting the value-added estimates employing empirical Bayes shrinkage to account for the estimation error. A conditional Bayes estimator is employed that results in the variance of these estimates equaling our best estimate of the variance in the actual value added of teachers (Carlin and Louis 1996). Statistics reported in this brief are for these conditional Bayes estimates.