Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment

Thomas S. Dee
Benjamin J. Keys

Abstract

A common criticism of merit-pay plans is that they fail to systematically target rewards to the most effective teachers. This study presents new evidence on this issue by evaluating data from Tennessee's Career Ladder Evaluation System and the Project STAR class-size experiment. Because the students and teachers participating in the experiment were randomly assigned, inferences about the relative quality of teachers certified by the career ladder should be unbiased. The results indicate that Tennessee's career ladder had mixed success in rewarding teachers who increased student achievement. Assignment to career-ladder teachers increased mathematics scores by roughly 3 percentile points but generally had smaller and statistically insignificant effects on reading scores. © 2004 by the Association for Public Policy Analysis and Management.

INTRODUCTION

The dramatic effects that teachers have on student achievement are largely undisputed. However, the effects associated with specific and measurable teacher inputs (e.g., experience, education, and certification) are hotly contested. One of the most important flash points for these disagreements involves the design of teachers' compensation. Teachers' salaries in the United States are typically based on a fixed schedule that only takes into account their years of experience and their education level. This "single salary" approach was widely adopted in the first half of the 20th century, partly as a response to the capriciousness and outright discrimination that had existed under more discretionary forms of compensation (Cohn, 1996; Odden and Kelley, 2002). But, in recent years, this fixed approach has been widely criticized for failing to attract, motivate, and retain high-quality teachers. In response, reforms that somehow link teachers' pay to performance have proliferated over the last 20 years. However, these "merit-pay" programs have typically been abandoned or sharply limited in scope after just a few years. Proponents of performance-based pay for teachers argue that the failure of merit pay is largely due to teacher and union opposition (e.g., Ballou, 2001). Critics counter that these failures reflect instead the fundamental technical difficulties involved in consistently identifying effective teachers and rewarding good teaching practices (e.g., Murnane and Cohen, 1986).

Manuscript received February 2003; review complete May 2003; revision complete August 2003; revision review complete November 2003; second revision complete December 2003; accepted February 2004

Journal of Policy Analysis and Management, Vol. 23, No. 3, 471–488 (2004)
© 2004 by the Association for Public Policy Analysis and Management
Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20022

Though there is widespread pessimism about whether merit-pay plans can solve the "evaluation problem" and reward good teaching, there is relatively little direct empirical evidence. And the evidence that does exist is actually quite positive, suggesting that students with merit-pay teachers do have higher gain scores (e.g., Cohn and Teel, 1991; Cooper and Cohn, 1997). However, the inferences based on conventional data may be contaminated by the non-random sorting of teachers and students that occurs both across and within schools. This study presents new evidence on whether a merit-pay system can systematically reward the teachers who are relatively effective at promoting measured achievement. This evidence is based on the unintended (and fortuitous) overlap between Tennessee's former merit-pay plan, the Career Ladder Evaluation System, and the well-known class-size experiment, Project STAR (Student Teacher Achievement Ratio).

There are two distinct reasons that the evaluations based on these data provide a powerful test of whether merit-pay systems can effectively reward good teachers. First, Tennessee's career-ladder program was well funded and was widely considered to have one of the nation's most sophisticated approaches to teachers' evaluation.[1] Specifically, Tennessee's career ladder combined multi-dimensional evaluations with large financial and professional rewards intended to provide stronger incentives for teachers' effort and to promote the recruitment and retention of high-quality teachers. Second, the availability of the contemporaneous data from the Project STAR class-size experiment provides an unusual opportunity to circumvent the confounding biases that can be generated in non-experimental settings. More specifically, as part of the class-size experiment, students and teachers within the participating schools were randomly assigned to their class types. These putatively random assignments imply that students' exposure to career-ladder teachers should be unrelated to their unobserved propensity for achievement.

Our test-score evaluations suggest that Tennessee's career ladder had only mixed success in targeting rewards to the more meritorious teachers. Specifically, our evidence indicates that assignment to a career-ladder teacher significantly increased mathematics scores by roughly 3 percentile points. However, most career-ladder teachers were not significantly more effective at promoting reading achievement. Furthermore, assignment to a teacher who had advanced further up the career ladder was not uniformly associated with significantly higher achievement. We also find that these inferences are not compromised by the violations of Project STAR's experimental design (i.e., classroom reassignment and attrition).

MERIT PAY FOR TEACHERS

The last 20 years have witnessed a sharply increased interest in educational reform and the proliferation of a variety of new policies at all levels of government. One of the earliest and most widely adopted types of educational reform over this period has involved merit pay for teachers. Under current practice, teachers typically begin with a base salary that increases only with their years of experience and additional education (Cohn, 1996; Odden and Kelley, 2002). This uniform pay structure has been widely criticized for failing to recognize the prevailing market demand for teachers with certain expertise (e.g., math and science), for failing to reward teachers' performance and productivity, and, ultimately, for discouraging high-ability

[1] One commentator praised Tennessee's career ladder as "perhaps the country's most comprehensive experiment in summative evaluation" (Brandt, 1995). However, some teachers in Tennessee alleged that the evaluation procedures were inadequate (Marquandt, 1995; May, 1990).

individuals from entering and remaining in a teaching career. Merit pay, which is defined broadly here as any system of teachers' compensation that explicitly rewards better performance, has been perceived as a potential remedy for the flaws of the single-salary system (e.g., Ballou and Podgursky, 2001; Hoerr, 1998). By 1986, 29 states had initiated some sort of merit pay for teachers (Cohn and Teel, 1992). However, these programs typically faced strong opposition from teachers' unions and were often abandoned or substantively altered after just a few years. Ballou and Podgursky (1997) report that only 12 percent of school districts use merit pay and that the amount of incentive pay in these districts averages only 2 percent of base pay.

One of the main objections to merit pay for teachers concerns the validity of the evaluation procedures and performance incentives (e.g., Elam, 1989). In an influential article, Murnane and Cohen (1986) note that, under an efficient merit-pay plan, employers should be able to explain clearly why an employee did not receive merit pay and what he would need to do to get it. Murnane and Cohen (1986) argue that these conditions are not met in the teaching profession where there is no single "blueprint" for effective practice. This "evaluation problem" is further complicated by the fact that schools have goals other than cognitive achievement (e.g., promoting citizenship, fostering individual development, and reducing drug use and violence) that are difficult to measure and often only achieved jointly through teacher cooperation. In other words, according to this line of reasoning, the diverse nature of educational outputs and the "imprecise" nature of effective teaching imply that it is infeasible to reward the isolated contributions of individual teachers. These concerns also suggest that the capricious results of most attempts to reward meritorious teachers could have perverse consequences. Merit-pay systems may distort the incentives for a variety of relevant teacher behaviors (e.g., cooperation, effort, and retention), as well as foster a demoralizing and unproductive work environment. Murnane and Cohen (1986) argue that these problems explain why merit-pay plans have often been dismantled. However, Ballou (2001) counters that merit pay is widely and successfully used in private schools, which suggests that there is nothing unique about the production of education that makes merit pay infeasible or unattractive.[2]

Pessimism about whether merit-pay systems can effectively reward good teachers is widespread. However, few studies have provided direct empirical evidence. And the evidence that is available has been surprisingly positive. For example, in regression analyses of class-level data from South Carolina, both Cohn and Teel (1992) and Cooper and Cohn (1997) find that classes taught by merit-pay teachers had significantly higher gain scores in math and reading. Similarly, related literature suggests that mathematics students learn more when their teachers have certification in mathematics (Darling-Hammond, Berry, and Thoreson, 2001; Goldhaber and Brewer, 2000, 2001; Wayne and Youngs, 2003). However, evidence of this sort should be interpreted with caution when making policy inferences. The critical concern is that the associations identified in these non-experimental data could reflect the confounding influence of other omitted variables. For example, these results are also consistent with the plausible hypothesis that the teachers who receive merit pay tend to select schools and classes whose students have unobserved propensities for high achievement (e.g., better socioeconomic priors, higher resource levels, increased parental involvement). If this were so, then regression results based on

[2] Ballou (2001) also notes that the amount of merit pay in private schools is quite large and attributes the frequent dismantling of alternative compensation for public school teachers to union opposition.

conventional data would overstate the success of the merit-pay programs in rewarding effective teachers. However, it should also be noted that the bias in a conventional evaluation could, quite plausibly, be in the other direction. For example, if principals assigned high-quality, merit-pay teachers to classes with unobserved propensities for low achievement, then a conventional regression analysis would understate the success of the merit-pay program.

TENNESSEE'S CAREER LADDER EVALUTION SYSTEM

This study presents new empirical evidence on whether merit-pay programs are successful in rewarding relatively effective teachers. This evidence is based on Tennessee's former merit-pay program, the Career Ladder Evaluation System, and data from a contemporaneous class-size experiment, Project STAR.[3] The availability of data from the Project STAR class-size experiment presents an unusual opportunity to eliminate the potential biases in inferences based on conventional data. Since the students and teachers within participating schools were randomly assigned, the within-school variation in the career-ladder status of teachers should be un-correlated with their students' unobserved propensities for achievement.[4] And an evaluation of Tennessee's merit-pay system may be particularly informative since the design of the program was considered unusually sound. The Career Ladder Evaluation System blended salary rewards with non-pecuniary benefits such as increased professional responsibilities (e.g., supervising beginning teachers, curriculum development). This emphasis on teachers' professionalism, as opposed to simply providing some teachers with more money for doing the same work, may have limited the scope for teachers' resentment, low morale, and uncooperative behavior. And the fact that Tennessee's career ladder did not establish quotas on the number of teachers who could receive awards also implies that it was less likely to harm teacher morale.[5] The career ladder also attempted to address teachers' concerns about the fairness of the assessments by relying on several data sources and evaluation instruments. Brandt (1995) praised Tennessee's program as "perhaps the country's most comprehensive experiment in summative evaluation." All of these design features imply that Tennessee's Career Ladder provides a particularly powerful test of whether merit pay can be effective in public schools.

Tennessee's Career Ladder Evaluation System was one of the most visible and widely discussed components of the broad reforms enacted by then-Governor Lamar Alexander as part of the state's Comprehensive Educational Reform Act (CERA) of 1984 (French, 1984). This merit-pay program was a form of "differentiated staffing" that combined a hierarchy of professional development (i.e., a career ladder) with financial and other professional rewards.

The career ladder consisted of five distinct stages. Fast-track options allowed those who had been teaching prior to CERA to advance to a career level subject to experience requirements and successful evaluations. However, for new teachers, the first rung of the career ladder was a 1-year "probation" supervised by two tenured teachers from their local school. Subject to a favorable review by their local school

[3] Project STAR lasted 4 years, beginning with kindergarten students in the fall of 1985.

[4] Evidence on the validity of the random within-school assignments is presented. Furthermore, this study adopts two approaches to dealing with the potentially confounding influence of behavioral responses to the experimental assignments (Krueger, 1999).

[5] In case studies of 18 school districts, Hartry, Greiner, and Ashford (1994) find that quotas were a major source of the morale problems associated with merit-pay plans.

district using state-approved criteria, these teachers were then placed on "apprentice status" for 3 years. At the end of those 3 years, local school districts could recommend that the teacher be granted a 5-year certification for "professional" or career level I status, which included a $1000 salary supplement from the state.

At the end of their 5-year certification, level I teachers could apply for another 5-year certification at level I. This level I re-certification was conducted by local districts, but was subject to oversight by a state official. A level I teacher could also seek a 5-year certification as a level II teacher. This advance required evidence of superior performance as defined by a state commission and the state board of education. Achieving level II status implied a $2000 state supplement for teachers choosing a 10-month contract and $4000 for those choosing an 11-month contract. At the end of a level II certificate, teachers could seek re-certification or advance to a 5-year certification as a level III teacher, which required further state-level evaluations but included state salary supplements of as much as $7000.

Evaluations at each stage of the career ladder assessed teachers on multiple domains of competence, using several distinct data sources (Furtwengler, 1985). However, on the first three rungs of the ladder (probation/apprentice, level I), the local school districts were primarily responsible for evaluating and certifying performance (Malo and French, 1987). The key evaluator at these stages—typically, the principal—received three to five days of state training on evaluation instruments and procedures. In contrast, the evaluations for certifications at levels II and III were largely conducted by a three-member team of peers from outside the teacher's school district. These evaluators, who received three to four weeks of training, were often level III teachers from other school districts who had been reassigned for a year by the state certification commission. The extensive training provided to those evaluating levels II and III teachers was considered appropriate since they fielded more complex evaluation instruments intended to discriminate among good, superior, and outstanding teachers (Malo and French, 1987).

Under the original formulation of the career ladder, participation was optional for veteran teachers, and mandatory for new teachers. It was initially expected that new teachers who failed to advance to level I status after their apprenticeship would be fired since they would no longer be eligible for the state portion of their salary (Belsie, 1984); but, in 1987, the career ladder was revised to make it optional for all teachers (Locker, 1991). Thus, the major consequence of failing to advance to level I status was essentially the lost opportunity for the new salary supplement. Relatively few teachers who participated in the career ladder appear to have faced this cost. Nearly all teachers (94 percent, according to Allison [1987]) chose to enter the career-ladder program. A 1991 state audit revealed that 95 percent of eligible teachers had achieved level I certification, prompting criticism that the standards for this designation had been severely diluted (Locker, 1991). However, only 79 percent of teachers applying for certification at levels II and III succeeded (Allison, 1987).

Though most teachers chose to participate and the success rates for certification were quite high, some teachers' strident criticisms echoed the issues Murnane and Cohen (1986) raised. For example, one fundamental criticism was that three classroom visits (some prearranged) were inadequate for evaluating teaching performance objectively (May, 1990). Some teachers also complained that separating the staff into levels strained relations and impaired morale (May, 1990). The application process was also criticized as an overly burdensome one that stressed "cunning and endurance … rather than merit" (Marquand, 1985). These comments suggest that, despite the relative sophistication of the career ladder, its efficacy in rewarding high-quality teachers is an open, empirical question. Furthermore, the high pass rates on

some of the career-ladder evaluations raise the possibility that, to the extent career-ladder teachers are relatively effective, it merely reflects the willingness of higher-quality teachers to apply, not the discriminating power of the evaluations.

DATA AND SPECIFICATIONS

Project STAR

Our evaluations provide evidence on the success of the career-ladder evaluations by analyzing data from Tennessee's Project STAR (Student Teacher Achievement Ratio). Project STAR is a well-known class-size experiment that began in the fall of 1985 with 6325 kindergarten students from 79 participating schools.[6] The experiment lasted 4 years (i.e., through the third grade). Overall, roughly 11,600 students participated with about 2200, 1600, and 1200 entering in the first, second, and third grades, respectively (Krueger, 1999).[7] Participating schools were drawn from around the state and, by legislative mandate, included inner-city and suburban schools from larger metropolitan areas (e.g., Knoxville, Nashville, Memphis, and Chattanooga), as well as rural schools and urban schools from smaller towns. The key feature of the experimental design was that students and teachers within participating schools and grades were randomly assigned to one of three class types: small classes, regular-sized classes, and regular-sized classes with teacher aides.[8] The putatively random within-school pairing of students with teachers implies that this experiment offers a unique and unintended opportunity to assess the relative quality of teachers who had been certified by the Career Ladder Evaluation System. More specifically, because a teacher's career-ladder status should be un-correlated with their students' unobserved propensity for achievement, the experiment may result in less biased measures of the association between career-ladder certification and true teacher quality.[9]

The evaluations presented here are based on student-level information from each study year, which were drawn from the Project STAR Public Access Data file. Pooling the student observations across all four grades implies a data set with roughly 24,000 observations (Table 1). The achievement outcomes available in this data set are based on student scores from grade-specific Stanford Achievement Tests in math and reading. More specifically, following Krueger (1999), the test outcomes modeled here are the grade and subject-specific percentile ranks based on these scores (Table 1).[10] The other available student-level variables include binary indicators for gender, race, age (here represented by a binary indicator for a birth year prior to 1980), as well as an indicator for whether the student received free lunches in their first study year (Table 1).[11] The public access data set also includes school

[6] For more detailed discussions of the Project STAR experiment in general, see Hanushek (1999), Krueger (1999), Mosteller (1995), and Word et al. (1990).
[7] The number of first-grade entrants was fairly high since kindergarten was not required.
[8] Because participating schools had to be large enough to accommodate one of each class type, smaller schools were excluded from the experiment.
[9] However, as with any social experiment, there are several ways in which behavioral responses might have confounded the experimental design and, by implication, the inferences we hope to make (Hanushek, 1999; Krueger, 1999).
[10] Test scores are missing for some students, largely due to absenteeism. For example, of the 6325 kindergarten students, test scores are available for only about 5900.
[11] A few observations were deleted because they lacked data on these traits. Given the very limited number of Hispanic, Asian, and American Indian Project STAR participants, our extract also includes only those observations from black and for white non-Hispanic students.

Table 1. Descriptive statistics, pooled K-3 Project STAR data.

Variable	Mean (Standard Deviation)
Student Traits	
Mathematics score	50.6
	(28.8)
Reading score	50.6
	(28.8)
Black	0.33
	(0.47)
Female	0.48
	(0.50)
Born before 1980	0.36
	(0.48)
Free lunch	0.49
	(0.50)
Small class assignment	0.30
	(0.46)
Teacher Traits	
Career ladder—	
probationary/apprentice	0.15
	(0.36)
Career ladder—level I	0.69
	(0.46)
Career ladder—level II or III	0.07
	(0.25)
Teacher experience	12.0
	(8.3)
Graduate degree	0.38
	(0.48)
Own-race teacher	0.78
	(0.42)

The mean mathematics score is based on 23,883 observations, the mean reading score mean on 23,544 observations. All other variables are based on 23,956 observations for which either test score is available.

identifiers, information on each student's class-type assignment in addition to information on participating teachers' race, years of experience, education, and career-ladder status.[12] Class-type assignment is represented here by a simple indicator for assignment to a small class since the prior empirical evidence indicates that teachers' aides were not an effective addition to regular-sized classes (e.g., Krueger, 1999). Teachers' race is represented by a binary indicator for whether the student has teacher of his or her own race since the relevance of a teacher's race appears to interact with the student's race (Dee, forthcoming). Each teacher's education level is represented by a binary indicator for whether she has a graduate degree. In the

[12] Notably, teacher gender is not included on the public-use data. In all likelihood, this was intended to preserve confidentiality since almost none of the teachers are male. Krueger (1999) reports that none of the kindergarten or first-grade teachers are male while 1 and 3 percent of the second-grade and third-grade teachers are, respectively.

initial test-score evaluations, each teacher's career-ladder status is represented by a binary indicator for whether he is in the career-ladder program. However, in some specifications, a less-restrictive approach is adopted in which career-ladder status is identified by binary indicators for three categories: a new entrant (i.e., apprentice or probationary status), a professional teacher (i.e., level I status), and a master teacher (i.e., level II or III status). The relevant reference category consists of teachers who were not on the career ladder (roughly 9 percent).[13] While this reference group may include some teachers who were denied access to the career ladder, the success rate of level I applicants suggests that it consists almost entirely of those who chose not to apply.

Specifications

The basic econometric model presented here relates Y_{isgc}, the grade and subject-specific percentile test rank for student i from school s, grade g, and class c, to student, teacher, and classroom traits and fixed effects for the grade, entry wave (kindergarten, grades 1 through 3), and the school of entry. More specifically, this model takes the following basic form:

$$Y_{isgc} = Z_{isgc}\Pi + X_{sgc}\beta + \alpha_g + \alpha_{sf} + \varepsilon_{isgc}$$

where α_g represents grade fixed effects, α_{sf} represents fixed effects for school-of-entry and entry-wave interactions and ε_{isgc} is a mean-zero random error.[14] The fixed effects for entry schools and entry waves are interacted because randomization occurred in the school of entry upon the year of entry (Krueger and Whitmore, 2001). And, since there is class-specific variation in class size and other unobserved determinants, class-specific heteroscedasticity in ε_{isgc} is accommodated in this model through the use of Huber-White standard errors. The matrix, **Z,** includes the variables that vary at the individual level (i.e., race, gender, age, and free lunch status). The matrix, **X,** includes class-specific variables, such as the teacher's career-ladder status and assignment to a small class. In the long form of this model, additional controls include years of teaching experience, the square of the experience measure, and binary indicators for whether the teacher has a graduate degree and whether the student has an own-race teacher. The robustness of the results to the inclusion of these additional variables is important since career-ladder status could easily proxy for the effects of these observed teacher traits.[15] Furthermore, since education and experience are the key components of the single-salary schedule, the results of the full model can indicate whether the career-ladder system improved upon those observables in rewarding effective teachers.

Because this is a somewhat unconventional regression analysis, there are several relatively unique issues of interpretation to note. First, one reasonable source of

[13] This includes 33 kindergarten students for whom the teacher's career-ladder status was listed as pending. The results are similar when these observations are excluded.

[14] The public-use data set did not include class identifiers. However, class identifiers were constructed by concatenating the available information on school, grade, class type, and teacher traits. These class identifiers were validated by replicating the class-size distribution reported by Krueger (1999).

[15] Because career-ladder status varies fairly closely with experience levels, it is possible that our results reflect effects related to teacher experience, which were not captured by the linear and quadratic experience terms. To address this possibility, we also estimated models in which years of teaching experience was interacted with dummy variables for whether teacher experience was in the typical probationary/apprentice range (0 to 3 years), in the level I range (4 to 8 years) or in the level II or III range (9 or more years). The empirical results were consistent with those reported based on these specifications.

concern is that our basic inferences might be confounded by the violations of Project STAR's experimental design. Student attrition from the schools participating in Project STAR was fairly high, ranging from 20 to 30 percent annually (Hanushek, 1999). Furthermore, the classroom assignments of some children who remained within the experiment were sometimes changed ("treatment crossover") in response to parental complaints and behavioral problems (Krueger, 1999). The possibly nonrandom nature of these class reassignments and attrition could lead to a flawed inference about the association between career-ladder status and effective teaching. Most obviously, if students with unobserved propensities for high achievement sought out master teachers through class reassignment or attrition, our basic results would overstate the quality of career-ladder teachers. Our expectation is that these concerns would have limited empirical relevance in this context. When choosing a particular school, parents presumably had fairly accurate expectations of the school's distribution of teacher quality. Furthermore, unlike a multi-year assignment to a particular class size, a 1-year assignment to a particular teacher does not provide a strong incentive for attrition or reassignment. However, we address these issues explicitly by assessing the randomness of the student-teacher assignments and by presenting the results from test-score models that correct for treatment crossover and attrition (Krueger, 1999).

A second and somewhat more subtle interpretative issue involves exactly what the regression estimates imply about the effects of the career-ladder program. The estimates should, of course, be narrowly understood as identifying the relative test score gains associated with assignment to a career-ladder teacher. However, to the extent career-ladder teachers are associated with higher student achievement, we should be cautious about exactly why this is so. The most conventional interpretation would be that the career ladder provided effective incentives for teacher effort and that the evaluations carefully discriminated among high- and low-quality teachers. However, the high pass rates on career-ladder evaluations suggest that these assessments were not particularly discriminating (at least at level I). This raises the possibility that, if career-ladder teachers were more effective, it is simply because higher-quality teachers were more willing to negotiate the bureaucratic impediments to advancing on the career ladder. In other words, it could be that the career ladder merely acted as a device for sorting teachers who have private, unobservable information about their quality. We cannot dismiss this possibility. However, it should be noted that the existence of high pass rates is also consistent with evaluations that are discriminating and teachers that, by and large, know that as well as their relative quality. Furthermore, the relative effectiveness of teachers at the master levels should speak to this issue somewhat since the pass rates for these designations, which involved more rigorous state-level evaluations, were lower. But, most important, it should be noted that these issues would not fundamentally complicate what our test-score results indicate about the basic efficacy of the career-ladder program. More specifically, if the career ladder led only to self-sorting of teachers by quality, it would still indicate that the program successfully directed its financial and professional rewards to meritorious teachers.

A third issue is that our estimates of the effects associated with probationary/apprentice status may be biased by the effects associated with all younger teachers, particularly since new teachers were required to enter the career ladder during the first 2 years of the Project STAR experiment. Where relevant, we assess the empirical relevance of this issue in several ways, including assessing the distinct test-score effects associated with the least experienced teachers. A fourth concern is that our interpretation would be biased if the career ladder influenced the

behavior of our reference group, the teachers who were not part of it. For example, the relative gains associated with career-ladder teachers could conceivably reflect a decrease in the quality of the reference teachers due to declines in morale and cooperative behavior among teachers (Murnane and Cohen, 1986). Alternatively, our results would understate the quality of career-ladder teachers (and the program, in general) if the reference teachers responded to the perceived challenge of the career ladder with increased effort (a "John Henry" effect, Cook and Campbell [1979]). While we cannot dismiss these possibilities, we consider them unlikely. In part, this is because we find that career-ladder teachers were relatively successful at promoting math, but not reading, achievement. It seems somewhat implausible that reform-driven changes in the behavior of teachers off the career ladder would significantly influence math, but not reading, achievement. Furthermore, the fairly high success rates of teachers choosing to enter or advance on the career ladder seems unlikely to trigger attrition, widespread resentment, or a sense of challenge among the teachers off of the career ladder. This is particularly so because the absolute level of financial well-being among teachers not on the career ladder was unaffected by the reforms. However, a potentially more relevant caveat that should be stressed is that, because our results only compare the relative performance among teachers, they do not identify the school-level effects that the career ladder could have had on teacher quality through influencing overall morale, cooperation, and teacher practices.

RESULTS

Table 2 presents the key results from OLS models for math and reading scores. The estimated coefficients on the student-level traits in these regressions uniformly indicate that test performance is significantly higher among whites, females, and students who are not eligible for free or reduced-price lunches. Younger students also had significantly higher performance on the reading tests. Interestingly, all eight of these regression models suggest that assignment to a career-ladder teacher led to increased test scores. Specifically, these results indicate that students with career-ladder teachers had math scores that were nearly 3 percentile points higher than those of students with other teachers. Similarly, these estimates suggest that reading scores were nearly 2 percentage points higher among these students. While it is difficult to assess the policy relevance of these changes, a simple comparison with the effects of other variables suggests that the effect sizes are quite large. For example, the estimated gains associated with assignment to a career-ladder teacher equal 40 to 60 percent of the estimated gains associated with assignment to a small class. Furthermore, these gains are approximately equivalent to a third of the corresponding black versus white gap in test scores. However, only the estimated effects on math scores are statistically distinguishable from zero.

The statistical insignificance of the reading results underscores an important caveat to this study. Because Project STAR was not designed to study the effects of career-ladder teachers, its statistical power to make meaningful inferences about even large effect sizes is relatively weak. In particular, our inferences may have weak power since we exploit the within-school variation in career-ladder status, which may be limited by the non-random sorting of teachers across schools. By way of contrast, the design of the experiment ensured that there was within-school variation in small class assignments. Correspondingly, the standard error on the small-class variable is roughly half the size of the standard error on the career-ladder variable. If the career-ladder variable had a similarly small standard error in the model for reading scores, the point estimate would be statistically significant.

Table 2. OLS estimates of the change in math and reading scores associated with assignment to a career ladder teacher.

	Math				Reading			
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Career ladder teacher	2.9†	2.7†	—	—	1.8	1.7	—	—
	(1.2)	(1.2)			(1.2)	(1.2)		
Probationary/ apprentice	—	—	2.9†	4.1†	—	—	0.13	2.3
			(1.5)	(1.6)			(1.41)	(1.5)
Level I	—	—	3.0†	2.4*	—	—	2.0*	1.5
			(1.2)	(1.3)			(1.2)	(1.2)
Level II or III	—	—	2.8	1.5	—	—	4.9‡	3.5†
			(1.8)	(1.8)			(1.7)	(1.7)
Teacher experience	—	0.17	—	0.30*	—	0.4‡	—	0.4‡
		(0.12)		(0.16)		(0.1)		(0.1)
Teacher experience squared	—	–0.005	—	–0.008†	—	–0.01‡	—	–0.01‡
		(0.004)		(0.004)		(0.003)		(0.003)
Graduate degree	—	0.6	—	0.8	—	0.1	—	–0.04
		(0.7)		(0.7)		(0.7)		(0.7)
Own-race teacher	—	3.8‡	—	3.8‡	—	3.0‡	—	3.0‡
		(1.0)		(1.0)		(0.9)		(0.9)
Small class assignment	—	4.5‡	—	4.6‡	—	4.5‡	—	4.4‡
		(0.7)		(0.7)		(0.6)		(0.6)
Female	1.1‡	1.1‡	1.1‡	1.1‡	5.5‡	5.5‡	5.5‡	5.5‡
	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)
Black	–11.0‡	–8.7‡	–11.0‡	–8.6‡	–6.7‡	–4.9‡	–6.7‡	–4.9‡
	(0.7)	(0.9)	(0.7)	(0.9)	(0.7)	(0.9)	(0.7)	(0.9)
Born before 1980	–0.2	–0.2	–0.2	–0.2	–2.6‡	–2.6‡	–2.6‡	–2.6‡
	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)
Free lunch	–10.7‡	–10.7‡	–10.7‡	–10.7‡	–11.8‡	–11.7‡	–11.8‡	–11.7‡
	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)
R^2	0.2137	0.2206	0.2137	0.2209	0.2354	0.2353	0.2364	0.2434
p-value (H_o: $\beta_1=\beta_2=\beta_3$)	—	—	0.9926	0.3665	—	—	0.0056	0.2081

Heteroscedastic-consistent standard errors are reported in parentheses. All models include grade fixed effects and the interactions of fixed effects for the entry school and entry wave. The *p* values refer to tests of the null hypothesis that the three career ladder coefficients are equal.
* Statistically significant at 10 percent level.
† Statistically significant at 5 percent level.
‡ Statistically significant at 1 percent level.

Table 2 also contains the key results from models that have a less-restrictive representation of career-ladder status. The reference group consists again of those teachers who are not on the career ladder. But three binary indicators are used to represent the "rung" of the students' teachers on the career ladder: probationary/apprentice status, level I and level II or III. The results from these models indicate that the test score changes associated with career-ladder status vary by subject. For example, the results for math scores indicate that it was the career-ladder teachers at the probationary/apprentice level and at level I who were particularly successful at promoting mathematics achievement. In contrast, career-ladder

teachers at the master level did not have a statistically significant effect on math scores. The imprecision of these estimates qualifies this evidence of heterogeneity. For example, the *p* values in the bottom row of Table 2 indicate that we cannot reject the hypothesis that the coefficients on these three career-ladder variables are equal. However, it is still noteworthy that point estimates associated with teachers at levels II and III are actually smaller than those for teachers at lower rungs.

This pattern could reflect the success of the career ladder in attracting (and retaining) new, high-ability math teachers or in providing teachers with early mentorship and professional development. However, as suggested earlier, this estimate may also be biased by the unobserved quality that could be unique to all younger teachers. More specifically, probationary and apprentice teachers may be particularly good at teaching math simply because newer teachers have higher average ability that declines as the cohort ages and as better teachers leave the profession. This is a plausible concern but we find that it is not supported by the data. First (and perhaps foremost), models that measure teacher experience with dummy variables (i.e., 0 to 3 years and 4 to 8 years) indicate that the least experienced teachers were actually relatively *ineffective* at promoting achievement in mathematics (though not significantly so). Second, the results in Table 2 indicate that this heterogeneity actually becomes noticeably stronger when we condition on teacher experience and its square. Third, the estimated effect associated with teachers at the probationary/apprentice level is similar in models that use the dummy variables for teacher experience. And, fourth, we also see a similar pattern of estimates when we only use students with teachers who have had five or more years of experience.[16]

Regardless, the pattern to the results in Table 2 suggests that, with regard to mathematics, the career ladder was not particularly effective at distinguishing good, superior, and outstanding teachers. In contrast, the reading results in Table 2 indicate that assignment to a level II or level III teacher was associated with a large and statistically significant increase in achievement. But the relative gains in reading achievement associated with teachers beginning the career ladder and those certified at level I were not statistically distinguishable from zero.

Threats to Internal Validity

The internal validity of the inferences presented in Table 2 turn in large part on the assumption that students and teachers were randomly assigned. However, as noted, that key assumption could have been effectively compromised by non-random attrition from the experiment as well as by treatment crossover. We examined these concerns in several ways. One basic approach to assessing the possible relevance of non-random reassignment or attrition is to consider whether the career-ladder status of teachers exhibits a persistent association with observable student traits that are associated with achievement. If non-random reassignment or attrition did confound the assignment to a teacher with a particular career-ladder status, we might expect to find that observed student traits (e.g., race, gender, age, free-lunch status, small class assignment) are correlated with the career-ladder status of their teachers. We present evidence on this by estimating auxiliary regressions where these student traits were the dependent variables and the independent variables were the three career-ladder dummy variables and the fixed effects for grade and for entry

[16] These last three specification checks rely on the fact that there are a non-trivial number of relatively experienced teachers who have probationary/apprentice status. These teachers may be relative latecomers to the career ladder as well as those for whom fast-track options were unavailable (Belsie, 1984).

Table 3. Tests of within-school differences in career-ladder status.

Student Trait	*p* value
Black	0.1350
Female	0.4281
Born before 1980	0.2691
Free lunch	0.3186
Small class assignment	0.9761

Each *p* value is for an *F* test of the null hypothesis that the three career-ladder variables have no effect on the given student-level variable. All models include grade fixed effects and the interactions of fixed effects for the entry school and entry wave.

school by entry wave. Table 3 presents the *p* values for tests of the joint significance of the three career ladder in each of these regressions. These results uniformly suggest that the career-ladder status of the teachers was unrelated to these important student traits.

Nonetheless, we also assess the empirical relevance of treatment crossover and student attrition in two other ways (Krueger, 1999). First, we examine the potentially confounding influence of non-random class reassignment by treating the career-ladder variables as endogenous regressors and generating two-stage least squares (2SLS) estimates of their association with test scores. We use as instrumental variables for these regressors the intended career-ladder status of each student's teacher (i.e., the "intent to treat"). More specifically, we constructed a measure for the intended career-ladder status of each student's teacher by identifying the career-ladder status of the teacher each student would have had if he or she had remained with their initial entry school and their initial class-size type. Because some schools had multiple classrooms with the same class-size type, this approach does not unambiguously identify the career-ladder status of the teacher each student would have had if they remained in their initial class assignment. However, in these limited instances, the instrumental variables simply take on fractional values for the mean probability of getting a teacher with a given career-ladder status in a given school and class type.[17]

We present the key results from these 2SLS models in Table 4. These specifications are similar to the long-form models in Table 2 and include the teacher and small-class controls. The 2SLS results are generally quite similar to the OLS results in Table 2, suggesting that treatment crossover did not confound our basic results. For example, these results indicate that assignment to a career-ladder teacher had large, positive effects on student achievement, increasing math scores by 3.9 percentile points and reading scores by 2.1 percentile points. And, as with the OLS results, only the implied change in math scores is statistically distinguishable from zero. While both of these 2SLS estimates are noticeably larger than their corresponding OLS estimates, these differences are not large relative to the sampling variation.[18]

The 2SLS estimates of the relative effects of teachers on different rungs of the career ladder are also similar to the OLS results in Table 2. Specifically, they suggest that there were large and statistically significant gains in math scores among

[17] In the few cases where the class type or grade for an entry school was not observed, students were assigned the mean value for the school and grade or for the school.

[18] In particular, Hausman tests indicate that we cannot reject the null hypothesis that the OLS and 2SLS estimates are equal.

Table 4. 2SLS estimates of the change in math and reading scores associated with assignment to a career-ladder teacher.

	Math		Reading	
Variable	(1)	(2)	(3)	(4)
Career-ladder teacher	3.9‡	—	2.1	—
	(1.5)		(1.4)	
Probationary/apprentice	—	8.1‡	—	2.9
		(2.2)		(2.1)
Level I	—	2.7*	—	1.7
		(1.6)		(1.5)
Level II or III	—	1.4	—	5.2†
		(2.2)		(2.1)
R^2	0.2205	0.2198	0.2431	0.2432
p value (H_o: $\beta_{OLS} = \beta_{2SLS}$)	0.2364	0.0363	0.7323	0.0743
p value (H_o: $\beta_1 = \beta_2 = \beta_3$)	—	0.0262	—	0.0482

Heteroscedastic-consistent standard errors are reported in parentheses. All models include grade-fixed effects and the interactions of fixed effects for the entry school and entry wave. Each model also includes the controls for student's race, age, gender, and socioeconomic status as well as the controls for small-class assignment and the teachers' traits. The first row of p values is from Hausman tests comparing the OLS and IV estimates. The second row of p values is from tests of the null hypothesis that the career-ladder coefficients are equal.
* Statistically significant at 10 percent level.
† Statistically significant at 5 percent level.
‡ Statistically significant at 1 percent level.

students assigned to teachers at the beginning of the career ladder. And students assigned to teachers certified at levels II and III experienced large and statistically significant gains in reading scores. While this pattern of results parallels the OLS results, the 2SLS point estimates are again noticeably larger and statistically meaningful, particularly for the math gains associated with assignment to a probationary/apprentice teacher. In these models, the hypothesis that the career-ladder variables have effects of the same size can be rejected. Hausman tests suggest that the equivalence of the OLS and 2SLS estimates can also be rejected. Interestingly, the implied bias in the OLS estimates suggests that parents or principals moved students with a propensity for high achievement away from teachers at the lowest rung of the career ladder.

While the existence of treatment crossover did not fundamentally confound our career-ladder results, our inferences could be biased by non-random attrition from the Project STAR experiment. For example, the apparent success of career-ladder teachers in promoting achievement could simply reflect the possibility that students with an unobserved propensity to achieve were also more likely to leave the experiment when assigned to a teacher who was not on the career ladder. We address this possibility by evaluating test score equations that include imputed test scores for students who left the experiment. More specifically, test score outcomes were crudely imputed for students who were absent or left the experiment by relying on the prior and subsequent subject-specific test score rankings available in the data set. In the models that include these imputed test scores, the key independent variables reflect the career-ladder status of the teacher the student would have had if they had remained with their entry school and classroom type (i.e., the "intent to treat" variables used as instruments in the 2SLS models). Since we do not know the

exact classroom some those who left the experiment would have attended, heteroscedasticity is accommodated at the level of the unique cells defined by the interactions of entry school, grade and classroom type.

The key results from these reduced-form evaluations are presented in Table 5. As a point of reference, the left panel of Table 5 presents the reduced-form estimates when only the actual test scores are used. The results in the right panel are based on actual test scores as well as imputed scores for students who left the experiment. The point estimates from models that include the imputed data tend to be smaller, which suggests that students with an unobserved propensity for low achievement may have been more likely to leave the experiment when assigned to a career-ladder teacher. However, these estimates are still quite similar to the OLS results in Table 2 and indicate that attrition from Project STAR does not fundamentally confound our career-ladder results. Specifically, these estimates suggest that assignment to a career-ladder teacher increased math scores by a statistically significant 2.7 percentile points and reading scores by a weakly significant 1.6 percentile points. Interestingly, the weak significance of the reading results in these models clearly reflects the gain in statistical precision associated with the increased sample size. As with the OLS and 2SLS results, these models also indicate that students experienced significant math gains when assigned to teachers on the lower rungs of the career ladder and significant reading gains when assigned to teachers at the top of the career ladder. However, the hypothesis that teachers on different rungs of the career ladder have the same effect cannot be rejected.

Table 5. Reduced-form estimates of the change in math and reading scores associated with intended career-ladder status.

Variable	Actual Test Scores		Actual and Imputed Test Scores	
	Math	Reading	Math	Reading
Intended career ladder teacher	3.7‡	2.0	2.7‡	1.6*
	(1.3)	(1.3)	(0.9)	(0.9)
R^2	0.2187	0.2403	0.2147	0.2360
Intended career ladder—probationary/apprentice	5.3‡	0.8	3.6‡	0.8
	(1.7)	(1.7)	(1.2)	(1.2)
Intended career ladder—level I	3.4†	2.1	2.6‡	1.7*
	(1.3)	(1.3)	(0.9)	(0.9)
Intended career ladder—level II or III	2.9	5.6‡	1.7	3.4†
	(2.0)	(1.9)	(1.4)	(1.4)
R^2	0.2190	0.2411	0.2148	0.2362
p value (H_o: $\beta_1 = \beta_2 = \beta_3$)	0.2997	0.0216	0.3602	0.1305
Sample size	23,883	23,544	34,317	33,978

Heteroscedastic-consistent standard errors are reported in parentheses. All models include grade fixed effects and the interactions of fixed effects for the entry school and entry wave. Each model also includes the controls for student's race, age, gender, socioeconomic status, and intended class type. The p values are from tests of the null hypothesis that the career ladder coefficients are equal.

* Statistically significant at 10 percent level.

† Statistically significant at 5 percent level.

‡ Statistically significant at 1 percent level.

CONCLUSIONS

The No Child Left Behind Act of 2001 requires that every teacher of core content in a public school be "highly qualified" by the end of the 2005–2006 school year. However, states appear to have considerable latitude in how they determine whether their teachers meet this Federal standard (Keller, 2004). As states grapple with the problems of designing policies that identify (and perhaps reward) qualified teachers, the debate that has surrounded merit pay is likely to be revisited. Over the last 20 years, merit-pay reforms have been widely implemented at the state and district level. But, typically, these reforms only lasted for a few years. The conventional interpretation of these early reforms has been that their failure reflects the inappropriateness of merit pay in educational settings. In particular, the critics of performance-based pay argue that merit-pay reforms fail in part because of the many difficulties involved in accurately assessing teacher quality. However, the proponents of merit pay dispute this assertion, pointing to the frequent use of merit pay in private schools. They also argue that the demise of these merit-pay reforms is often due to the opposition of teacher unions.

Surprisingly, there has been relatively little direct empirical evidence on whether merit-pay plans have been successful in targeting rewards to effective teachers. This study presented new empirical evidence on this issue by examining data from Tennessee's Project STAR class-size experiment and the contemporaneous Career Ladder Evaluation System. Overall, the results suggest that this career-ladder system was at least partially successful at rewarding teachers who were relatively effective at promoting student achievement. Specifically, this evidence indicates that assignment to a teacher who had been certified by the career-ladder evaluations led to large and statistically significant increases in mathematics scores (roughly 3 percentile points). However, these gains in mathematics scores appear to have been somewhat concentrated among teachers who were on the lower rungs of the career ladder. In contrast, only assignment to a teacher who had reached the top of the career ladder led to statistically significant gains in reading achievement. One important caveat to these results is that, since Project STAR was not designed to study the effects of career-ladder teachers, our models may merely have weak statistical power for making meaningful inferences about some of the smaller effects. It should also be noted that, because our study only compared the relative performance of teachers, our results do not speak to the issue of how the existence of the career ladder may have, positively or negatively, influenced the school-level teaching environment.

The implications of these results for the desirability of other merit-pay programs are decidedly mixed. The qualified successes of Tennessee's program clearly suggest the possibility that teacher quality can be reliably rewarded when there is a well-designed evaluation system. Furthermore, the availability of data from an experiment in which students and teachers were randomly assigned allows us to be unusually certain that our inferences were not confounded by the non-random sorting of students and teachers. However, the evidence that teachers on higher rungs of the career ladder were not uniformly better also underscores the considerable challenge of designing a system of teacher compensation that rewards quality in a fair and equitable manner.

We thank three anonymous referees for their useful comments, and the Smith Richardson Foundation for its financial support. The usual caveats apply.

THOMAS S. DEE is an Assistant Professor in the Department of Economics at Swarthmore College.

BENJAMIN J. KEYS is a graduate student in the Department of Economics at the University of Michigan.

REFERENCES

Allison, S. (1987). Teachers qualify for more pay. United Press International, January 30.

Ballou, D. (2001). Pay for performance in public and private schools. Economics of Education Review, 20(1), 51–61.

Ballou, D., & Podgursky, M. (1997). Teacher pay and teacher quality. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Ballou, D., & Podgursky, M. (2001). Defining merit: Let the market decide. Education Matters, 1(1), 16–25.

Belsie, L. (1984). Tennessee turns to merit pay to help attract teachers. Christian Science Monitor, May 8, p. 3.

Brandt, R. M. (1990). Incentive pay and career ladders for today's teachers. Albany: State University of New York Press.

Brandt, R. M. (1995). Teacher evaluation for career ladder and incentive pay programs. In D.L. Duke (ed.), Teacher evaluation policy: From accountability to professional development. Albany: State University of New York Press.

Cohn, E. (1996). Methods of teacher remuneration: Merit pay and career ladders. In W.E. Becker & W.J. Baumol (eds.), Assessing educational practices: The contribution of economics (pp. 209–238). Cambridge and London: MIT Press; New York: Russell Sage Foundation.

Cohn, E., & Teel, S. J. (1992). Participation in teacher incentive program and student achievement in reading and math. Proceedings of the Business and Economic Statistics Section, American Statistical Association.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton Mifflin.

Cooper, S. T., & Cohn, E. (1997). Estimation of a frontier production function for the South Carolina educational process. Economics of Education Review, 16(3), 313–327.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. Educational Evaluation and Policy Analysis, 23(1), 57–77.

Dee, T. S. (in press). Teachers, race and student achievement in a randomized experiment. Review of Economics and Statistics.

Elam, S. M. (1989). The second Gallup/Phi Delta Kappan poll of teachers' attitudes towards the public schools. Phi Delta Kappan, 70(10), 785–798.

French, R. L. (1984). Dispelling the myths about Tennessee's career ladder program. Educational Leadership, 41(4), 9–13.

Furtwengler, C. (1985). Tennessee's career ladder plan: They said it couldn't be done! Educational Leadership, 43(3), 50–55.

Goldhaber, D. D., & Brewer, D. J. (2000). High school teacher certification status and student achievement. Educational Evaluation and Policy Analysis, 22(2), 129–145.

Goldhaber, D. D., & Brewer, D. J. (2001). Evaluating the evidence on teacher certification: A rejoinder. Educational Evaluation and Policy Analysis, 23(1), 79–86.

Hanushek, E. A. (1999). 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(2), 143–164.

Hatry, H. P., Greiner, J. M., & Ashford, B. G. (1994). Issues and case studies in teacher incentive plans (2nd Edition). Washington, DC: Urban Institute Press.

Hoerr, T. R. (1998). A case for merit pay. Phi Delta Kappan, 80(4), 326–327.

Keller, B. (2004). Rigor disputed in standards for teachers. Education Week, January 14.

Krueger, A. B. (1999). Experimental estimates of education production functions. Quarterly Journal of Economics, 114(2), 497–532.

Krueger, A. B., & Whitmore, D. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project STAR. Economic Journal, 111(468), 1–21.

Locker, R. (1991). Career ladder weak at base, audit shows; 95% of teachers reach lowest rung. The Commercial Appeal (Memphis, TN), July 11, p. A1.

Malo, G. E., & French, R. L. (1987). The Tennessee career ladder: What it is and how it has changed. Thresholds in Education, 13(1), 16–19.

Marquand, R. (1985). A lesson in school reform from Tennessee. Christian Science Monitor, September 9, p. 28.

May, L. (1990). Teachers dislike merit pay program. Los Angeles Times, December 22, p. A4.

Mosteller, F. (1995). The Tennessee study of class size in early grades. Critical Issues for Children and Youth, 5, 113–127.

Murnane, R. J., & Cohen, D. K. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. Harvard Educational Review, 56(1), 1–17.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. Review of Educational Research, 73(1), 89–122.

Word, E., Johnston, J., Bain, H. P., Fulton, B. D., Zaharias, J. B., Achilles, C. M., Lintz, M. N., Folger, J., & Breda, C. (1990). The state of Tennessee's Student/Teacher Achievement Ratio (STAR) project final summary report, 1985–1990. Nashville: Tennessee State Department of Education.