

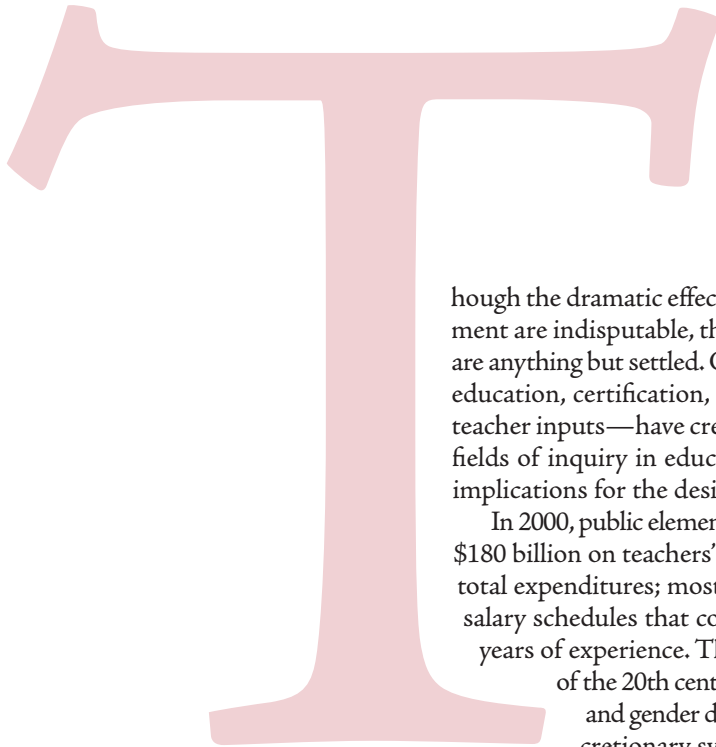
Dollars

What a
Tennessee
experiment
tells us
about
merit pay



ILLUSTRATION FROM GETTY IMAGES

AND Sense



hough the dramatic effects that teachers have on student achievement are indisputable, the exact ingredients of effective teaching are anything but settled. Questions about how to value experience, education, certification, and pedagogical skills—the big four of teacher inputs—have created one of the most highly contentious fields of inquiry in education, particularly since they have clear implications for the design of teacher compensation systems.

In 2000, public elementary and secondary schools spent roughly \$180 billion on teachers' salaries and benefits, about half of their total expenditures; most of it was distributed according to fixed salary schedules that considered only a teacher's education and years of experience. This system has its origins in the first half of the 20th century and was partly a response to the racial and gender discrimination that existed under more discretionary systems at that time.

However, over the past 20 years more educators have wondered whether such pay packages can attract, motivate, and retain high-quality teachers in a highly competitive professional world (see *Forum*, page 8). In response to such concerns, there was a flurry of merit pay activity in the early 1980s. Twenty-nine states had initiated some sort of merit pay program for teachers by 1986. Since then, however, almost all of them have been diluted or discontinued. A 1997 study by economists Dale Ballou and Michael Podgursky reported that 12 percent of school districts were using merit pay in some way, but the amount of incentive in these districts averaged only two percent of base pay.

Critics of merit pay argue that the falloff in such programs was due to the fundamental technical difficulties of accurately identifying effective teachers and rewarding good teaching practices. Proponents of performance-based pay insist that these experiments were too limited in scope and were destined to fail in the face of stiff opposition from teachers and unions.

BY THOMAS S. DEE AND BENJAMIN J. KEYS

Despite widespread pessimism among educators about whether merit pay systems can effectively reward good teachers, most of the limited empirical evidence has been surprisingly positive.

Despite widespread pessimism among educators about whether merit pay systems can effectively reward good teachers, most of the limited empirical evidence has been surprisingly positive. For example, two studies (in 1992 and 1997) found that the math and reading test scores of students in South Carolina improved significantly when the students were taught by teachers receiving merit pay. Similarly, related and more recent literature suggests that mathematics students learn more when their teachers have certification in mathematics.

Policymakers should be cautious in interpreting this sort of evidence, however, as the apparent benefits of certification could merely reflect differences in the students placed in their classrooms. For example, teachers who receive merit pay may tend to select schools and classes whose students are high achievers for other reasons. Likewise, parents especially engaged in their children's education may work to ensure their assignment to teachers with strong credentials. Such subtle differences may not be visible in the data typically available to researchers.

New Evidence

The effectiveness of these short-lived merit pay programs is exceptionally difficult to measure because of these selection effects. However, the fortuitous overlap of two Tennessee programs from the mid-1980s and 1990s provides an unusual opportunity to circumvent this problem. Project STAR (Student Teacher Achievement Ratio) was a large-scale class-size experiment that began with kindergarten students in the fall of 1985. At roughly the same time, the Volunteer State began directing pay increases to teachers deemed meritorious under a Career Ladder Evaluation System.

The fact that both teachers and students in schools participating in Project STAR were assigned randomly to classrooms allows for an especially rigorous test of whether a merit pay system can effectively reward good teachers.

Before describing the Tennessee programs in detail, however, we need to take a closer look at some of the objections to merit pay for teachers. One concerns the problem of designing valid evaluation procedures for measuring teacher performance. Under an

efficient merit pay plan in any industry, employers should be able to explain clearly why an employee did not receive merit pay and what he or she would need to do to get it. Whether these conditions can be met in the teaching profession, where there is no single blueprint for effective practice, has been the most contentious issue surrounding merit pay. This evaluation problem is further complicated by the fact that schools have goals other than cognitive achievement (for instance, promoting citizenship, fostering individual development, and reducing drug

use and violence) that are difficult to measure and are often achieved only with teachers' cooperation.

These concerns raise the possibility that attempts to reward meritorious teachers could even have perverse consequences. For example, merit pay systems may discourage cooperation among teachers or otherwise foster a demoralizing and unproductive work environment.

While these problems may explain why merit pay plans have often been dismantled, some researchers suggest that they are excuses, not reasons. Dale Ballou, an economist at Vanderbilt University, has argued that merit pay is widely and successfully used in private schools, which suggests that there is nothing unique about education that makes merit pay infeasible or unattractive. Ballou notes that the amount of merit pay in private schools is quite large and that the teachers who report receiving it have earnings that are nearly 10 percent higher than their nonmerit counterparts. In contrast, the earnings of merit pay teachers in public schools are only 2 percent higher than their nonmerit colleagues. Ballou attributes the frequent dismantling of alternative compensation for public school teachers to union opposition.

The Career Ladder

Can we devise a merit pay system that overcomes the challenges of definitional clarity and valid measurement? Can we do so without directly incorporating measures of students' progress on standardized tests?

The Tennessee programs initiated by Governor Lamar Alexander in 1984 offer some reason to believe that we can. At the same time, they underscore the considerable difficulty of doing so in a fair, equitable, and effective manner.



Part of Tennessee's Comprehensive Educational Reform Act, the Career Ladder Evaluation System was both well funded and sophisticated in its approach to teacher evaluation. As Richard M. Brandt, a professor at the University of Virginia, wrote in 1995, the program was "perhaps the country's most comprehensive experiment in summative evaluation."

Governor Alexander, who would go on to become secretary of education under President George H. W. Bush and is now a U.S. senator, was more colloquial in his description, calling the program "an old-fashioned horse trade with teachers. Taxpayers said to teachers, 'The state will pay you up to 70 percent more based on your performance if you'll promise to be evaluated every five years.'"

Rung by Rung

While it lasted—for 13 years—the now-defunct Career Ladder had many of the elements that merit pay backers believed a good program should have, including multidimensional evaluations and a hierarchy of professional development (in other words, a career ladder) that was coordinated with significant financial and professional rewards.

The ladder had five distinct stages, ranging from probationary to master. Fast-track options allowed those who had been teaching before 1984 to advance immediately, subject to successful evaluations, to a career level matching their experience.

For new teachers, however, the first rung of the career ladder was a one-year probation supervised by two tenured teachers from their school. Subject to a favorable review by the school district, using state-approved criteria, these teachers were then placed on apprentice status for three years. At the end of those three years, the school district could recommend that the teacher be granted a five-year certification for professional, or Career Level I, status, which included a \$1,000 salary supplement from the state.

Then, at the end of the five-year Level I stage, a teacher could either apply for another five-year Level I certification or seek a five-year certification as a Level II teacher. Advancement required evidence of superior performance, as defined by a state commission and the state board of education, but it also came with a \$2,000

Governor Alexander called the program "an old-fashioned horse trade with teachers. Taxpayers said to teachers, 'The state will pay you up to 70 percent more based on your performance if you'll promise to be evaluated every five years.'"

state supplement for those who chose a 10-month contract and \$4,000 for those choosing an 11-month contract, a significant bonus to teachers' salaries at the time.

At the end of the Level II certification period, the same kind of option was available: a teacher could seek recertification at Level II or pass more rigorous evaluations to receive a Level III certification and a salary supplement of as much as \$7,000.

The evaluations that occurred at each stage of the career ladder assessed teachers on multiple "domains of competence" using several distinct data sources (such as student and principal questionnaires, peer evaluations, a teacher's portfolio, and a written test). On the first three rungs of the ladder (probation, apprentice, Level I), the local school districts were responsible for evaluating and certifying performance. 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 conducted largely by a three-member team of peers from outside the teacher's district. These evaluators received three to four weeks of training and were often Level III teachers from other districts who had been borrowed

for a year by the state certification commission. The extensive training provided to the Level II and Level III evaluators was considered appropriate since they fielded more complex evaluation instruments intended to discriminate among "good, superior, and outstanding" teachers.

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. However, in 1987, the career ladder was revised to make it optional for all teachers. The major consequence of failing to advance to Level I status was essentially the lost opportunity for the salary supplement.

Interestingly, it appears that relatively few teachers faced this cost. Nearly all of the state's teachers (94 percent of them, according to one report) chose to enter the career-ladder program. A state audit in 1991 revealed that 95 percent of eligible teachers had achieved



National Writing Board

The National Writing Board, founded in **1998**, has now given an independent, unbiased assessment, with two Readers, of high school history papers from **29** states, and sent each author a three-page report, with scores and comments, which she/he has asked us to send to college admissions officers (at **70** colleges so far), or simply could use as feedback on one of her/his best history research papers. History research papers of two lengths—around 2,000 words, or around 5,000 words—with (Turabian) endnotes and bibliography, may be submitted, with a notarized Submission Form and a check for \$100, (we spend more than three hours on each paper), made out to the National Writing Board, to: the National Writing Board, 730 Boston Post Road, Suite 24, Sudbury, Massachusetts 01776. Deadlines are November 1 and June 1 each year. The following **32** colleges and universities now endorse this independent assessment service for academic writing: [for more information: www.tcr.org; fitzhugh@tcr.org]

Amherst
Bowdoin
Carnegie Mellon
Claremont McKenna
Colgate
Connecticut College
Dartmouth
Duke
Eckerd
Emory
Georgetown
Hamilton
Harvard
Haverford
Illinois Wesleyan
Lafayette

Lehigh
Michigan
Middlebury
Northwestern
Notre Dame
Pitzer
Princeton
Richmond
Sarah Lawrence
Spelman
Trinity (CT)
Tufts
University of Virginia
Washington and Lee
Williams
Yale

University of Virginia

Office of Admission

August 2, 2004

Will Fitzhugh, President
National Writing Board
730 Boston Post Road, Suite 24
Sudbury, Massachusetts 01776

Dear Mr. Fitzhugh,

The University of Virginia is pleased to endorse the **National Writing Board** and the important project you have undertaken. Nurturing and enhancing the experience and skills in writing and doing research are among the most important challenges for our country and I commend you for founding and leading this effort.

We would be happy to place your brochures in our reception room so that more high school students will consider doing it.

Sincerely,
John A. Blackburn
Dean of Admission

Level I certification, prompting criticism that the standards for this designation had been severely diluted. However, among teachers applying for certification at Levels II and III, the success rate was only 79 percent.

Though most teachers chose to participate, and the success rates for certification were quite high, some expressed criticisms that echoed issues often raised by merit pay critics: for example, that three classroom visits (some of them prearranged) were inadequate for evaluating teaching performance objectively and that separating the staff into levels strained relations among teachers and hurt morale. Even the application process was criticized for emphasizing, as the *Christian Science Monitor* reported, “cunning and endurance . . . rather than merit.” The criticisms suggest that, despite the relative sophistication of the career ladder, its efficacy in rewarding high-quality teachers remains an open question.

Project STAR

Coincidentally, a compelling way to evaluate the success of the career ladder system comes via data from Governor Alexander’s Student Teacher Achievement Ratio program. Project STAR was an experimental study of class-size reduction that also began in the fall of 1985. That year, it included 6,325 kindergarten students from 79 participating schools. The experiment lasted for three more years, following students through the 3rd grade. Overall, roughly 11,600 students participated, with additional students entering the participating schools in the 1st, 2nd, and 3rd grades. Participating schools were drawn from around the state and, by legislative mandate, included inner-city and suburban schools from larger metropolitan areas (Knoxville, Nashville, Memphis, and Chattanooga) as well as rural schools and those from smaller towns. All students in classrooms included in the experiment were given the Stanford Achievement Tests in math and reading in the spring of each year.

Pooling the information from the experiment’s four years yields a single data set with roughly 24,000 student observations for each subject. Roughly one-third of these observations are for black students, and nearly half were for students eligible for the free-lunch program. Fully 91 percent of the student observations in the dataset come from classrooms taught by teachers participating in the career ladder: 15 percent had teachers with probationary or apprentice status, 69 percent had teachers at Level I, while just seven percent had teachers who had reached Level II or III.

The criticisms suggest that, despite the relative sophistication of the career ladder, its efficacy in rewarding high-quality teachers remains an open question.

The key feature of the experimental design of Project STAR was that students and teachers within participating schools and grades were randomly assigned to one of three class types: small classes, regular-sized classes, or regular-sized classes with teacher aides. These random assignments allow us to use the STAR data to compare the performance of students assigned to career-ladder teachers with the performance of students in the same school and grade who were assigned to nonparticipating teachers.

Restricting the comparison to students attending the same school is essential because student-teacher pairings were random only within a given school. That is, the experiment did not move students and teachers to schools they would not otherwise have attended or staffed. This unfortunately means that some schools in the data set—those with classrooms taught by teachers with the same career-ladder status—do not offer useful information for looking at the effects of career-ladder status.

It should also be noted that student attrition from schools participating in the experiment was high, ranging from 20 to 30 percent each year, and that roughly 10 percent of students moved between small and regular classes. While most of the movement between classes was due to parental complaints or behavioral problems, the attrition figures could also reflect other factors unrelated to the study, such as students’ moving out of a school’s geographic zone or having to repeat a grade. However, if parents of students with unobserved propensities for high achievement sought out master teachers by class reassignment or by moving to another school altogether, our results would overstate the quality of career-ladder teachers.

Fortunately, we expect that these problems are less important for a study of the career ladder than for one about class size. Unlike a multiyear assignment to a particular class size, a one-year assignment to a particular teacher does not provide a strong incentive for attrition or reassignment. Students would be assigned a new teacher in the next academic year. By contrast, students placed in a large class were expected to remain in large classes through the 3rd grade.

Still, to evaluate whether the experiment successfully matched students and teachers randomly within schools, we examined the association between students’ traits and their assignment to a teacher of their own race. If the pairings of students and teachers were indeed random and remained so over time, we should find no within-school association between observed student traits and exposure



to teachers in the career ladder. As expected, students' race, gender, age, eligibility for the free-lunch program, and class-size assignment all exhibit small and statistically weak within-school relationships with assignment to a career-ladder participant.

Finally, because the student-teacher pairings were initially random, any statistically significant difference in performance between students with and without career-ladder teachers should be attributable to true differences in the quality of the teachers. The most conventional interpretation of such performance differences would be that the program provided effective incentives for teachers and that the evaluations carefully discriminated among teachers of high and low quality. However, the high pass rates on career-ladder evaluations suggest that these assessments were not particularly discriminating (at least through Level I). This raises the possibility that, if career-ladder teachers were more effective, it was simply because better teachers were more willing to negotiate the bureaucratic impediments to advancing on the career ladder. Nonetheless, even if the career ladder led only to self-sorting of teachers by quality, it would indicate that the program successfully directed its financial and professional rewards to meritorious teachers.

Results

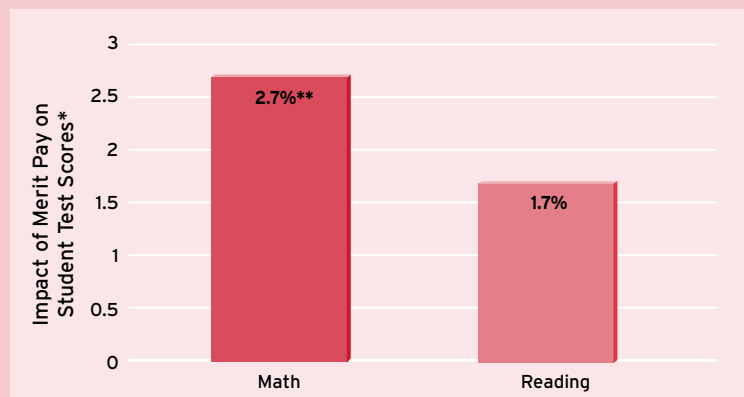
To see what these Tennessee programs tell us about merit pay, let's first look at the effects simply of having a teacher in the career-ladder program, ignoring for the moment the teacher's specific level of accomplishment. To eliminate the effects of any chance differences in performance caused by other observable characteristics, our analysis takes into account students' age, gender, race, and eligibility for the free lunch program; whether they had been assigned to a small class; and whether they were assigned to a teacher of the same race—which earlier research using these same data found to have a large positive effect on student performance (see "The Race Connection," Spring 2004). We also include as control variables two conventional indicators of teacher quality: experience and possession of a graduate degree.

Our main results indicate that students with career-ladder teachers scored nearly 3 percentile points higher in mathematics than students with other teachers. They also suggest that reading scores were nearly 2 percentile points higher among these students, though the results for reading fall just short of conventional levels of statistical significance (see Figure 1).

The estimated effects on reading scores are statistically indistinguishable from zero primarily because they are less precise. If the effect on reading performance of having a career-ladder teacher were as precisely estimated as the effect of being

Merit Pay for Teachers Equals Higher Scores for Students (Figure 1)

K–3 students whose teachers earned merit pay scored higher in both math and reading than those students whose teachers were paid under the conventional structure.



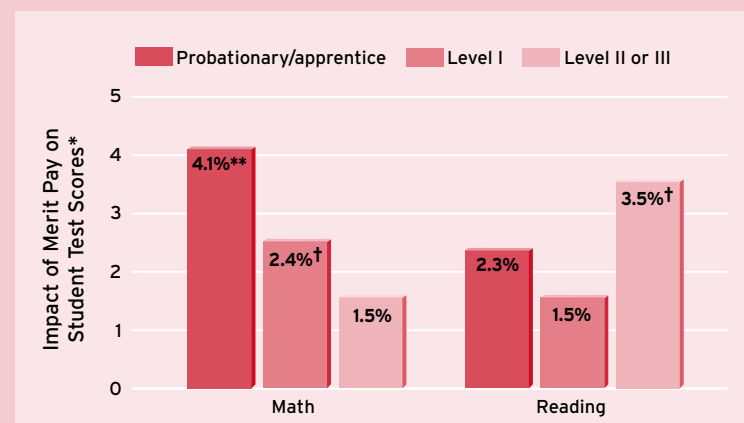
* Percentile scores

** Statistically significant at the 5 percent level

SOURCE: Authors' calculations

Merit Pay Works for Beginning Teachers — and Experienced Ones Too (Figure 2)

Reading scores of students with career teachers rose sharply when merit pay was introduced. Math gains were greatest for students of newer teachers.



* Percentile scores

** Statistically significant at the 5 percent level

† Statistically significant at the 10 percent level

SOURCE: Authors' calculations

in a smaller class, it would also be statistically significant. That it is not may reflect the fact that the experiment was designed to evaluate the effects of differences in class size, not the career-ladder program.

Regardless, our best guess is that having a career-ladder teacher in either subject had a quite large effect. The estimated gains associated with assignment to a career-ladder teacher equal 40 to 60 percent of the gains associated with assignment to a

Students with career-ladder teachers scored nearly 3 percentile points higher in mathematics than students with other teachers.

class with roughly 15 students rather than 22. Furthermore, the gains are approximately equivalent to a third of the black-white gap in test scores among students in the experiment.

When evaluating these results, it is important to keep in mind that 91 percent of the student observations in the data set came from classrooms with teachers certified by the career ladder. The benefits of having a career-ladder teacher are measured relative to a somewhat atypical base—namely, the small group of students whose teachers chose not to apply for the program or were unsuccessful in their application.

Our second analysis, therefore, considered not only the teacher's participation in a career ladder, but also the teacher's status within the program. That is, we looked separately at the effects of having a teacher at the probationary or apprentice level, at Level I, and at Level II or III.

In math, the career-ladder teachers at the probationary/apprentice level and at Level I were the most successful at promoting achievement. In contrast, career-ladder teachers at the master level did not have a statistically significant effect on math scores (see Figure 2).

This surprising pattern could in theory reflect the success of the career ladder in attracting (and retaining) new, high-ability math teachers and in providing these new teachers with early mentoring and professional development. However, an alternative explanation is that novice teachers, many of whom quickly leave teaching, happen to be particularly adept at teaching math. The fact that we have already controlled for differences in teachers' experience makes this explanation unlikely. Moreover, a similar pattern emerges when we look only at students with teachers having five or more years of experience, a good number of whom remained at the probationary/apprentice level (perhaps because fast-track options were not available in their area). In short, it appears that the career ladder simply was not very effective at distinguishing superior or outstanding math teachers from those who were merely competent.

In reading, by contrast, assignment to a Level II or Level III teacher was associated with a large and statistically significant increase in reading achievement, while estimates of the effects of having a teacher from both of the other two groups remained positive but statistically insignificant. This suggests the career ladder may have been modestly successful in identifying the most outstanding teachers in reading.

Conclusions

Overall, our results suggest that Tennessee's Career Ladder Evaluation System was at least partially successful at rewarding teachers who were relatively effective at promoting student achievement. Though the program was voluntary for veteran teachers, the combination of large bonuses and relatively undemanding evaluations—

at least at the lower levels—led the vast majority of teachers to enter. Nonetheless, assignment to a teacher who had been certified by the career-ladder evaluations led to large and statistically significant increases in mathematics scores and sizable, though statistically insignificant, increases in reading scores.

But our findings also suggest that the teachers who were on the highest rungs of the career ladder (and received the largest pay increases) were not consistently better at promoting student achievement. In reading, only students with a teacher at the highest levels of the career ladder made statistically significant gains. In contrast, the math-score gains associated with having a career-ladder teacher actually appear to have been concentrated among students with teachers on the *lowest* rungs of the career ladder. These mixed findings underscore the challenge of designing a system of teachers' compensation that rewards quality in a fair and equitable manner—a political challenge as much as a technical one.

Despite some success in rewarding teachers for producing better student outcomes, the career ladder was a target of the same criticisms that challenge virtually all attempts to tinker with systems of teachers' compensation. A few years of budgetary constraints helped kill the will to keep it all together. Thus, having made participation in the career ladder voluntary for teachers in 1987, it was perhaps inevitable that the Tennessee legislature in 1997 voted to prevent additional teachers from

entering the program and becoming eligible for merit bonuses. Teachers already in the program, though no longer subject to regular evaluations, were allowed to keep their bonuses for the duration of their careers.

As Lamar Alexander lamented at the time, "Those who questioned the Model-T Ford didn't try to kill it. They replaced it with something better." Continuing debates over merit pay programs in districts in Tennessee and beyond indicate that efforts to find such a replacement are under way. But it may still be too early to tell whether the future for merit pay for teachers will resemble that of the Edsel or the Mustang.



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.