



The Teacher Incentive Fund Pay-for-Performance Bonuses – Four-Year Follow-Up

Chiang, H., Speroni, C., Herrmann, M., Hallgren, K., Burkander, P., Wellington, A. (2017). Evaluation of the teacher incentive fund: final report on implementation and impacts of pay-for-performance across four years. *Institute of Education Sciences*, NCEE 2018-4004.

Overview:

Between 2006 and 2012, the U.S. Department of Education’s Teacher Incentive Fund (TIF) awarded \$1.8 billion in grants to school districts to support the establishment of performance-based compensation systems for teachers and principals in high-need schools. This was a very large, well-conducted RCT, with a sample of 138 elementary and middle schools in 10 districts, which evaluated the impact of pay-for-performance bonus systems implemented with TIF grants awarded in 2010. At the end of the fourth year of TIF implementation, the study found very small effects on student achievement that were not statistically significant and so may be the result of chance rather than true effects of the program. Specifically, students in treatment schools scored at the 37th percentile on state tests in reading, versus the 35th percentile for students in control schools; and at the 37th percentile in math, versus the 36th percentile for the control group. (These effects equate to 0.04 standard deviations in both reading and math.) The effects during the first three years of program implementation were similarly small, and reached statistical significance in some years but not others.¹

Description of the Program:

All schools in the study—*i.e.*, both treatment and control schools—implemented three components of a performance-based compensation system. Specifically, they: (1) measured the effectiveness of educators (teachers and principals) using students’ achievement growth and at least two observations of classroom or school practices; (2) offered pay opportunities for educators to take on additional responsibilities; and (3) targeted professional development based on educators’ performance on the above effectiveness measures. In addition, schools in the treatment group paid “substantial, differentiated, and challenging to earn” bonuses to educators based solely on the educators’ performance on the effectiveness measures. These treatment school bonuses amounted to an average of \$2,058 per teacher per year. By contrast, control group schools paid their educators an automatic 1 percent bonus each year (unrelated to performance on the effectiveness measures), which amounted to an average of \$470 per teacher per year. Thus, the study evaluated the impact of pay-for-performance bonuses for educators compared to automatic bonuses. The four program components were implemented somewhat heterogeneously across districts.

Study Design:

In 2011, 138 low-income, low-performing elementary and middle schools in 10 school districts were randomly assigned to a treatment group, which implemented the program, or a control group which did not.² Seven schools closed or dropped out of the study after its inception, resulting in a sample of 131

¹ This study measured the *average* effect of the schools’ pay-for-performance bonus systems. These systems varied considerably in their key features across the study sites, and it is possible that some were effective even if the average impact across all systems was close to zero. An alternative study design could have sought to identify effective systems within TIF, rather than measure TIF’s average impact.

² These 10 districts represent fewer than 7 percent of the 148 school districts that implemented TIF in the 2014-15 school year. On average, districts included in the evaluation were larger and more likely to be urban and located in the West, and in states with collective bargaining agreements than non-evaluation TIF districts.

schools (65 treatment and 66 control) that implemented TIF in the 2011-12, 2012-13, 2013-14, and 2014-15 school years. State test scores in reading and math were collected for approximately 39,000 students in school at the time of testing during the fourth year.

Key Findings:

At the end of the fourth year of TIF implementation, the study found that students in treatment schools scored marginally higher on state standardized tests than students in control schools in reading (37th percentile statewide for treatment group students versus 35th percentile for control group students) and math (37th percentile statewide versus 36th percentile).³ These effects did not reach statistical significance at conventional levels, and so may be the result of chance rather than true effects of the program (although the effect on reading came close to significance, $p=0.08$). The effects of TIF during the first three years of implementation were also very small, and reached statistical significance in some years but not others.⁴

Summary of Study Quality:

This was a well-conducted RCT of a large federal program as it is typically implemented in high-need public schools. The treatment and control schools were highly similar in key pre-program characteristics (e.g., demographics, student achievement). The study's analysis appropriately accounted for minor differences in pre-program math scores as well as for the fact that schools (not individual students) were randomly assigned to the treatment and control groups. Only 5 percent of the randomized schools dropped out of the study, and the rate of missing student test score data in the remaining schools was 8 percent in the fourth year and did not differ between the treatment and control group.

³ These effects equate to 0.04 standard deviations in reading and 0.04 standard deviations in math.

⁴ In addition to examining TIF's effects in the original sample of 138 schools, the study examined its effects in a second cohort of 45 schools randomized in the spring or summer of 2012. The treatment group schools in this cohort implemented TIF during the 2012-2013, 2013-2014, and 2014-2015 school years. Adding the new cohort of schools to the first cohort resulted in effects sizes during the three years after random assignment that were the same or slightly smaller than those seen in the first cohort alone.