What Works Clearinghouse™

ies INSTITUTE OF EDUCATION SCIENCES

September 2013

# WWC Review of the Report "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools"[1]

**The findings from this review do not reflect the full body of research evidence on the New York City Schoolwide Performance Bonus Program.**

## What is this study about?

The study examined whether offering schoolwide performance bonuses to teachers had an effect on student achievement or teacher retention in New York City public schools.[2]

Researchers analyzed data on students and teachers from 396 high-need public elementary, middle, and high schools from 2007–08 through 2009–10. Of these schools, 233 were randomly assigned to the intervention group and 163 to the comparison group.

The study estimated the effects of the bonus program by comparing the outcomes in schools that were offered participation in the program—even if they ultimately declined to participate—with the outcomes in schools that were not offered the opportunity to participate.

## What research question does this study answer?

The primary research question for this study is "what is the impact of the performance bonus program on student achievement and teacher retention?"

Because some of the schools that were eligible to participate in the bonus program did not ultimately participate, the study estimated both an "intent-to-treat" (ITT) estimate of the effect of being eligible to participate in the program, as well as a "treatment on the treated" estimate of the effect of participating in the bonus program. This review focuses on the ITT estimates.

### Features of New York City's Schoolwide Performance Bonus Program

As part of its accountability system, the New York City Department of Education set school-level goals for student academic performance and growth for each school. Each year, it awarded Progress Report card scores to schools based on student achievement on state English language arts and math exams (25%), yearly student progress (60%), and measures of the learning environment (15%).

The Schoolwide Performance Bonus Program provided performance bonuses to school staff based on their schools' Progress Reports.

- The program operated in high-need schools from 2007–08 through 2010–11, with schools randomly assigned to either an intervention or comparison group in 2007–08.
- If a school was randomly selected for the program, it had to secure votes in favor of program participation from 55% or more of its full-time union teachers in order to be eligible for bonuses.
- Participating schools could receive lump-sum payments of $3,000 per union teacher for reaching 100% of their school-level goals, or $1,500 per union teacher for meeting at least 75% of their goals.
- A four-member, school-level compensation committee decided in advance how to distribute payments among teachers and other staff.

## What did the study find?

Study authors reported that the bonus program had statistically significant negative impacts on middle school achievement in math (author-reported effect size of –0.05) and English language arts (effect size of –0.03). In addition, the authors reported a statistically significant difference of –4.4 percentage points in high school graduation rates, reflecting lower graduation rates among students in intervention schools.

The study found that the teacher performance bonus program had no statistically significant impacts on elementary school achievement or teacher retention.

WWC Rating

***The research described in this report meets WWC evidence standards without reservations***

***Strengths:*** This study is a well-executed randomized controlled trial.

## Appendix A: Study details

### Fryer, R. G. (2011). *Teacher incentives and student achievement: Evidence from New York City Public Schools* (NBER Working Paper No. 16850). Cambridge, MA: National Bureau of Economic Research.

**Setting**

The study was conducted in New York City public schools.

**Study sample**

Three hundred and ninety-six high-poverty schools were included in the random assignment process: 187 elementary schools, 82 middle schools, 39 K–8 schools, 73 high schools, one K–12 school, and 14 schools serving middle and high school students. Schools were randomly assigned to an intervention (233) or comparison (163) group. Once offered the opportunity to participate in the bonus program, 55% of an intervention school's full-time union teachers had to vote in favor of participation. Schools were included in the analysis as part of their original randomly assigned condition, regardless of whether they ultimately participated in the bonus program.

**Intervention group**

As part of its accountability system, the New York City Department of Education (NYCDOE) gave each school goals for student academic performance and growth as measured by state math and English language arts tests and, to a lesser extent, student attendance. The intervention consisted of paying schools lump-sum bonuses for meeting those goals: $3,000 per union teacher for meeting all of its goals and $1,500 per union teacher for meeting 75% of its goals. A four-member committee in each school decided how the lump sum would be distributed across eligible recipients (e.g., equally distributed or some other method) with the constraint that bonus distribution could not be based on seniority alone.

**Comparison group**

Comparison group schools were not offered the opportunity to participate in the bonus program and continued with business-as-usual.

**Outcomes and measurement**

The study examined school performance on state standardized tests (for elementary and middle school students), 4-year high school graduation rates, and teacher retention. Student outcome data came from student-level school district records, and teacher outcome data came from district human resources records. For school achievement analyses, outcomes across three years were included. Outcomes from the first two years of implementation were used for teacher retention analyses. For a more detailed description of these outcome measures, see Appendix B.

**Support for implementation**

Schools were provided information about the how the teacher incentive program worked, including requirements for a school-level decision-making process for determining how the lump-sum performance bonus would be distributed among school staff.

**Reason for review**

This study was identified for review by the WWC by receiving significant media attention.

## Appendix B: Outcome measures for each domain

| | |
|---|---|
| **Reading achievement** | |
| *New York State English Language Arts (ELA) Achievement Test* | The New York ELA achievement test was developed by McGraw-Hill and was administered to students in grades 3 through 8 in New York public schools. The test included both multiple choice and short response sections and assessed student achievement in three areas: information and understanding, literary response and expression, and critical analysis and evaluation. Scores were obtained from student-level administrative records from the school district. They were standardized by grade level and academic year to have a mean of zero and standard deviation of one. |
| **Mathematics achievement** | |
| *New York State Mathematics Achievement Test* | The New York mathematics achievement test was developed by McGraw-Hill and was administered to students in grades 3 through 8 in New York public schools. The test included items on number sense and operations, algebra, geometry, measurement, and statistics. Scores were obtained from student-level administrative records from the school district. They were standardized by grade level and academic year to have a mean of zero and standard deviation of one. |
| **High school graduation** | |
| *4-year high school graduation rates* | Administrative records from the school district were used to determine whether students graduated within 4 years of entering high school. |
| **Teacher retention** | |
| *School retention* | Teacher-level human resources records were used to determine whether someone teaching at an elementary, middle, or K–8 school in the study sample in 2007 was teaching in the same school in the following year. |
| **Teacher behavior** | |
| *Personal absences* | Teacher-level human resources records were used to determine the cumulative number of personal absences teachers took during the school year. |

**Table Notes:** Some outcomes were included in the study but are not included in this review. They include: high school Regents Exams scores (excluded because it was not possible to distinguish between first-time and repeat test takers); rate of 4-year graduation with Regents Diploma (excluded because this outcome was similar to the 4-year graduation rate); and teacher retention in the district (excluded because this outcome was similar to the school retention outcome and had some issues with its measurement). In addition, outcomes that were identified by the author as "alternative" were not included in this review (student attendance, student behavior problems, grade point average, and predictive ELA and math scores) because they were not the focus of the study.

## Appendix C: Study findings for each domain

| Domain and outcome measure | Study sample | Sample size | Mean (standard deviation) | | WWC calculations | | | *p*-value |
|---|---|---|---|---|---|---|---|---|
| | | | Intervention group | Comparison group | Mean difference | Effect size | Improvement index | |
| **Reading achievement** | | | | | | | | |
| *New York State ELA Achievement Test* | Elementary schools | 227 schools/ 175,894 students | –0.38 (0.88) | –0.36 (0.89) | –0.01 | –0.01 | 0 | > 0.05 |
| *New York State ELA Achievement Test* | Middle schools | 136 schools/ 147,141 students | –0.56 (0.79) | –0.53 (0.80) | –0.04 | –0.04 | –2 | < 0.05 |
| **Domain average for reading achievement** | | | | | | **–0.03** | **–2** | **Statistically significant** |
| **Mathematics achievement** | | | | | | | | |
| *New York State Math Achievement Test* | Elementary schools | 227 schools/ 176,387 students | –0.44 (0.89) | –0.42 (0.90) | –0.02 | –0.02 | –1 | > 0.05 |
| *New York State Math Achievement Test* | Middle schools | 136 schools/ 147,493 students | –0.58 (0.84) | –0.54 (0.86) | –0.06 | –0.06 | –2 | < 0.05 |
| **Domain average for mathematics achievement** | | | | | | **–0.04** | **–2** | **Statistically significant** |
| **High school graduation** | | | | | | | | |
| *4-year graduation rate* | High schools | 88 schools/ 27,995 students | 0.55 (0.50) | 0.58 (0.49) | –0.03 | –0.09 | –3 | < 0.05 |
| **Domain average for high school graduation** | | | | | | **–0.09** | **–3** | **Statistically significant** |
| **Teacher retention** | | | | | | | | |
| *School retention (Elementary)* | Elementary school teachers | 187 schools/ 21,700 teachers | 0.81 (0.39) | 0.82 (0.38) | –0.01 | –0.03 | –1 | > 0.05 |
| *School retention (Middle)* | Middle school teachers | 82 schools/ 8,289 teachers | 0.73 (0.44) | 0.76 (0.43) | –0.03 | –0.09 | –4 | > 0.05 |
| *School retention (K–8)* | K–8 school teachers | 39 schools/ 4,693 teachers | 0.79 (0.41) | 0.79 (0.41) | 0.00 | –0.01 | 0 | > 0.05 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Domain average for teacher retention** | | | | | | **–0.04** | **–2** | **Not statistically significant** |
| **Teacher behavior** | | | | | | | | |
| *Personal absences (Elementary)* | Elementary school teachers | 187 schools/ 18,543 teachers | 7.85 (7.99) | 7.57 (7.64) | 0.29 | 0.04 | +2 | > 0.05 |
| *Personal absences (Middle)* | Middle school teachers | 82 schools/ 6,727 teachers | 7.47 (7.72) | 7.91 (7.60) | –0.47 | –0.06 | –2 | > 0.05 |
| *Personal absences (K–8)* | K–8 school teachers | 39 schools/ 3,977 teachers | 8.03 (8.01) | 7.42 (8.53) | 0.60 | 0.07 | +3 | > 0.05 |
| **Domain average for teacher behavior** | | | | | | **+0.02** | **+1** | **Not statistically significant** |

**Table Notes:** For mean difference, effect size, and improvement index values reported in all domains except teacher behavior, a positive number favors the intervention group and a negative number favors the comparison group. For the teacher behavior domain, a negative number favors the intervention group and a positive number favors the comparison group, because greater teacher absences is considered a negative outcome. For student-level outcomes including math and ELA test scores and high school graduation rates, the effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average student's outcome that can be expected if the student is exposed to the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is exposed to the intervention. Similarly, for teacher retention outcomes, the effect size is a standardized measure of the effect of an intervention on teacher outcomes, representing the change (measured in standard deviations) in an average teacher's likelihood of retention that can be expected if the teacher is exposed to the intervention. The WWC-computed domain average effect size is a simple average rounded to two decimal places; the average improvement index is calculated from the average effect size. Effect sizes of teacher-level outcomes should not be compared with those of student-level outcomes because they are relative to a different distribution. The statistical significance of the study's domain average was determined by the WWC; the study is characterized as demonstrating potentially negative effects on high school graduation, reading achievement, and mathematics achievement because the univariate statistical tests for at least one measure in each domain are negative and statistically significant with the others statistically insignificant. The study is characterized as demonstrating indeterminate effects on teacher retention and teacher absences because none of the univariate statistical tests in the domain are statistically significant or substantively important. ELA = English language arts.

**Study Notes:** The means presented in this table are regression-adjusted, provided by the author at the request of WWC. For all continuous outcomes, adjusted means and raw standard deviations are reported in the table and effect sizes are computed using Hedges' *g*. For all binary outcomes, unadjusted means and standard deviations are reported to facilitate ease of interpretation and effect sizes are computed using Cox's index. WWC calculations differ slightly from author calculations. Mean differences in the table may not correspond directly to subtracting the comparison from the intervention means because of rounding errors. The effect sizes presented for the achievement outcomes (ELA and math tests) are based on regressions with controls using scores from all three implementation years and standardized to represent 1-year impacts. The authors controlled for student characteristics, including prior achievement test scores, and school characteristics. For teacher absences, teacher demographic variables and teacher value added estimates for ELA and math for the year before program implementation were used as controls. The *p*-values presented in the table were reported in the original study. The authors made corrections for clustering.

## Endnotes

[1] Single study reviews examine evidence published in a study (supplemented, if necessary, by information obtained directly from the author[s]) to assess whether the design meets WWC evidence standards. The review reports the WWC's assessment of whether the study meets WWC evidence standards and summarizes the study findings following WWC conventions for reporting evidence on effectiveness. This study was reviewed using the single study review protocol, version 2.0. The WWC rating applies only to the results that were eligible under this topic area and met WWC standards without reservations or met WWC standards with reservations, and not necessarily to all results presented in the study.

[2] Some outcomes were included in the study but are not included in this review. These include: high school Regents Exams scores (excluded because it was not possible to distinguish between first-time and repeat test takers); rate of 4-year graduation with Regents Diploma (excluded because this outcome was similar to the 4-year graduation rate); and teacher retention in the district (excluded because this outcome was similar to the school retention outcome and had some issues with its measurement). In addition, outcomes that were identified by the author as "alternative" were not included in this review (student attendance, student behavior problems, grade point average, and predictive ELA and math scores) because they were not the focus of the study.

## Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2013, September). *WWC review of the report: Teacher incentives and student achievement: Evidence from New York City Public Schools.* Retrieved from http://whatworks.ed.gov

## Glossary of Terms

**Attrition** Attrition occurs when an outcome variable is not available for all participants initially assigned to the intervention and comparison groups. The WWC considers the total attrition rate and the difference in attrition rates across groups within a study.

**Clustering adjustment** If intervention assignment is made at a cluster level and the analysis is conducted at the student level, the WWC will adjust the statistical significance to account for this mismatch, if necessary.

**Confounding factor** A confounding factor is a component of a study that is completely aligned with one of the study conditions, making it impossible to separate how much of the observed effect was due to the intervention and how much was due to the factor.

**Design** The design of a study is the method by which intervention and comparison groups were assigned.

**Domain** A domain is a group of closely related outcomes.

**Effect size** The effect size is a measure of the magnitude of an effect. The WWC uses a standardized measure to facilitate comparisons across studies and outcomes.

**Eligibility** A study is eligible for review if it falls within the scope of the review protocol and uses either an experimental or matched comparison group design.

**Equivalence** A demonstration that the analysis sample groups are similar on observed characteristics defined in the review area protocol.

**Improvement index** Along a percentile distribution of students, the improvement index represents the gain or loss of the average student due to the intervention. As the average student starts at the 50th percentile, the measure ranges from –50 to +50.

**Multiple comparison adjustment** When a study includes multiple outcomes or comparison groups, the WWC will adjust the statistical significance to account for the multiple comparisons, if necessary.

**Quasi-experimental design (QED)** A quasi-experimental design (QED) is a research design in which subjects are assigned to intervention and comparison groups through a process that is not random.

**Randomized controlled trial (RCT)** A randomized controlled trial (RCT) is an experiment in which investigators randomly assign eligible participants into intervention and comparison groups.

**Single-case design (SCD)** A research approach in which an outcome variable is measured repeatedly within and across different conditions that are defined by the presence or absence of an intervention.

**Standard deviation** The standard deviation of a measure shows how much variation exists across observations in the sample. A low standard deviation indicates that the observations in the sample tend to be very close to the mean; a high standard deviation indicates that the observations in the sample are spread out over a large range of values.

**Statistical significance** Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than 5% ($p < 0.05$).

**Substantively important** A substantively important finding is one that has an effect size of 0.25 or greater, regardless of statistical significance.

Please see the WWC Procedures and Standards Handbook (version 2.1) for additional details.