What Works Clearinghouse™

ies INSTITUTE OF EDUCATION SCIENCES

October 2012

WWC Review of the Report "Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program"[1]

The findings from this review do not reflect the full body of research evidence on the *NYC School-Wide Performance Bonus Program*.

What is this study about?

The study examined the effects of offering a schoolwide teacher performance bonus program on students' reading and mathematics achievement.[2]

The study sample included 309 high-poverty New York City public schools serving students in grades K–8 from 2007–08 to 2008–09. Of these schools, 181 were randomly chosen to be offered the opportunity to participate in the performance bonus program. The comparison group consisted of 128 schools that did not receive the chance to participate.

The study estimated the effects of the bonus program by comparing outcomes from the intervention group schools—even if they ultimately declined to participate in the bonus program—with the outcomes from the comparison group.

What research question does this study answer?

This study aims to address how performance-based bonus programs impact schoolwide achievement. Some of the schools that were eligible to participate in the bonus program did not ultimately participate. Therefore, study authors 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 the *NYC School-Wide Performance Bonus Program*

Schools that were given the opportunity to participate in this program were eligible to earn schoolwide bonuses for meeting school-level goals. School goals were tied to the district's accountability system, which awarded letter grades to schools based on student achievement on state reading and math exams, yearly student progress, and measures of the learning environment.

To accept the offer to participate in the performance bonus program, schools had to secure votes in favor of participation from 55% or more of their full-time union staff. Among the schools offered participation in the program, 25 (approximately 14%) voted not to participate or withdrew.

Participating schools could receive lump-sum payments of $3,000 per union teacher or $1,500 per union teacher for meeting at least 75% of their school-level goals. A compensation committee consisting of the principal, a second administrator, and two union representatives elected by the school's union members decided in advance how payments for reaching school-level goals would be distributed among teachers and other staff, within program guidelines.

What did the study find?

The study found that the offer of a schoolwide teacher performance bonus program did not have a statistically significant effect on students' reading achievement in either 2007–08 or 2008–09 or on mathematics achievement in 2007–08. For 2008–09, study authors reported a very small, but statistically significant, negative effect of the bonus program on mathematics achievement.

WWC Rating

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

Strengths: The study is a well-implemented randomized controlled trial.

Cautions: The changes in observed achievement may be in part due to (1) changes in student learning, (2) the movement of students of differing achievement levels between schools, or (3) a combination of both effects. This analysis cannot separate these effects; it can only report on their combined impact.

In addition, because the study analyzed school level effects, the magnitude of the effects reported cannot be directly compared to the magnitude of an effect of an intervention that uses student-level data for the analysis.

Appendix A: Study details

Goodman, S. F., & Turner, L. J. (2010). *Teacher incentive pay and educational outcomes: Evidence from the New York City Bonus Program*. New York: Columbia University.

Setting

The study was conducted in New York City.

Study sample

The study sample included schools that were randomly assigned to an intervention (181) or comparison (128) group in the first year of program implementation (309 schools total). In the second year of the study, four of the original intervention schools closed, bringing the total school count to 305. Schools were identified for the study according to the following selection process. First, to be eligible to participate in the schoolwide performance bonus program, schools had to serve students in grades K–8 and be designated as "high need." 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; this vote was held in November 2007. Of the 181 schools originally assigned to the intervention condition, 25 ultimately did not participate in the bonus program because they failed to reach this agreement threshold. In addition, two other intervention schools were moved to the comparison group before being notified of their study group assignment, and four schools originally assigned to the comparison group were allowed to vote and ultimately chose to participate in the bonus program. In all cases, schools were included in the analysis as part of their originally randomly assigned condition, regardless of participation status.

Intervention group

As part of its accountability system, the New York City (NYC) Department of Education gave each school goals for student academic performance and growth as measured by state mathematics and English language arts tests and, to a lesser extent, student attendance. At the end of the year, the system assigned each school a letter grade based on the extent to which it met the goals. The intervention consisted of paying schools lump-sum bonuses for meeting those goals: $3,000 per union teacher for meeting all its goals, and $1,500 per union teacher for meeting 75% of its goals. A four-member committee in each school decided ahead of time how the lump-sum bonus would be distributed across staff (e.g., equally distributed or some other method) with the constraints that all union teachers must receive a bonus payment and that seniority not be the only basis for differential payments to teachers.

Comparison group

The comparison group schools did not participate in the teacher performance bonus program but continued with business-as-usual. In the context of NYC, schools with "A" or "B" grades received rewards such as principal bonuses and additional funds when students transferred into the school from a school with a poor grade. Schools with a "D" or "F" grade faced consequences such as school closure or principal removal.

Outcomes and measurement

In both years of the study, student achievement outcomes were measured by the state standardized assessments in English language arts (administered in January) and mathematics (administered in March) in grades 3–8, aggregated to the school level.

Support for implementation

No support was offered to schools to implement this intervention.

Reason for review

This study was identified for review by receiving media attention.

Appendix B: Outcome measures for each domain

Reading achievement	
Schoolwide mean score on the New York State Assessment Program's English Language Arts (ELA) Test	The New York ELA achievement test was developed by McGraw-Hill and was administered to students in grades 3–8 in New York public schools. The ELA test included both multiple choice and short response sections. The test assessed student achievement in three areas: information and understanding, literary response and expression, and critical analysis and evaluation. The exam was administered in January of each school year.
Schoolwide percentage proficient on the New York State Assessment Program's ELA Test	The percentage of students in a school scoring proficient or better on the state ELA assessment.
Mathematics achievement	
Schoolwide mean score on the New York State Assessment Program's Mathematics Test	The New York mathematics achievement test was developed by McGraw-Hill and was administered to students in grades 3–8 in New York public schools. The mathematics test included items on number sense and operations, algebra, geometry, measurement, and statistics. The exam was administered in March of each school year.
Schoolwide percentage proficient on the New York State Assessment Program's Mathematics Test	The percentage of students in a school scoring proficient or better on the state mathematics assessment.

Table Notes: The study also examined teacher absences, teacher turnover, and the characteristics of newly-hired teachers using aggregated data on individual teachers. However, the manuscript lacked sufficient information on the data sources and aggregation methods to determine the validity of these measures. Therefore, these outcomes are not included in this single study review.

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								
Schoolwide mean score on the New York State Assessment Program's ELA Test—January 2008	Elementary and middle schools, Year 1	309 schools	nr	nr	–0.395	nr	nr	> 0.05
Schoolwide mean score on the New York State Assessment Program's ELA Test—January 2009	Elementary and middle schools, Year 2	305 schools	nr	nr	–0.584	nr	nr	> 0.05
Domain average for reading achievement						**nr**	**nr**	**Not statistically significant**
Mathematics achievement								
Schoolwide mean score on the New York State Assessment Program's Mathematics Test—January 2008	Elementary and middle schools, Year 1	309 schools	nr	nr	–0.789	nr	nr	> 0.05
Schoolwide mean score on the New York State Assessment Program's Mathematics Test—January 2009	Elementary and middle schools, Year 2	305 schools	nr	nr	–1.385	–0.05	–2	< 0.05
Domain average for mathematics achievement						**nr**	**nr**	**Statistically significant**

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on school outcomes, representing the change (measured in standard deviations) in a school's outcome that can be expected if the school receives the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in a school's percentile rank that can be expected if the school is given the intervention. The study is characterized as having no discernible effects on Reading achievement because there were no statistically significant or substantively important effects in that domain. The study is characterized as having potentially negative effects on Mathematics achievement because one of the estimated effects in that domain was negative and statistically significant and the other was not statistically significant. nr = not reported. ELA = English language arts.

Study Notes: No corrections for clustering or multiple comparisons were needed. The mean differences and p-values reported in this table are from Models 2 and 5 in Table 2 of the manuscript and represent the coefficient on the intervention variable in a regression of the outcome in question on an intervention indicator and a set of covariates. A corrected effect size for Mathematics, Year 2, was provided to the WWC by study authors and is reflected in the table. Authors did not report the effect sizes for the other outcomes examined.

Appendix D: Supplemental findings by 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 proficiency								
Schoolwide percent proficient on the New York State Assessment Program's ELA Test—January 2008	Elementary and middle schools	309 schools	nr	nr	–0.009	nr	nr	> 0.05
Schoolwide percent proficient on the New York State Assessment Program's ELA Test—January 2009	Elementary and middle schools	305 schools	nr	nr	–0.013	nr	nr	> 0.05
Mathematics proficiency								
Schoolwide percent proficient on the New York State Assessment Program's Mathematics Test—January 2008	Elementary and middle schools	309 schools	nr	nr	–0.009	nr	nr	> 0.05
Schoolwide percent proficient on the New York State Assessment Program's Mathematics Test—January 2009	Elementary and middle schools	305 schools	nr	nr	–0.017	nr	nr	< 0.05

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on school outcomes, representing the change (measured in standard deviations) in a school's outcome that can be expected if the school receives the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in a school's percentile rank that can be expected if the school is given the intervention. nr = not reported. ELA = English language arts.

Study Notes: No corrections for clustering or multiple comparisons were needed. The mean differences and *p*-values reported in this table are from Models 2 and 5 in Table 3 of the manuscript and represent the coefficient on the intervention variable in a regression of the outcome in question on an intervention indicator and a set of covariates. The authors did not report effect sizes for these outcomes.

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 study 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 review protocol and met WWC standards without reservations or met WWC standards with reservations, and not necessarily to all results presented in the study.

[2] The study also examined teacher absences, teacher turnover, and the characteristics of newly-hired teachers using aggregated data on individual teachers. However, the manuscript lacked sufficient information on the data sources and aggregation methods to determine the validity of these measures. Therefore, these outcomes are not included in this single study review.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2012, October). *WWC review of the report: Teacher incentive pay and educational outcomes: Evidence from the New York City Bonus Program.* 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.