What Works Clearinghouse™

ies INSTITUTE OF EDUCATION SCIENCES

September 2013

WWC Review of the Report "A Big Apple for Educators: New York City's Experiment with Schoolwide Performance Bonuses. Final Evaluation Report"[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 monetary bonuses for teachers improved schoolwide academic achievement in New York City public schools.

Study authors analyzed data from 389 high-need elementary, middle, and high schools in New York City in the first year of the bonus program (2007–08) and from 371 of those same schools in the second (2008–09) and third (2009–10) years. These schools had been randomly assigned to either an intervention or a comparison group in 2007–08.

The researchers assessed the effectiveness of the bonus program by comparing the scores on the New York City Department of Education's (NYCDOE) Progress Reports for schools randomly assigned to the intervention group with those of the comparison group.[2]

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 schoolwide achievement?"

The analysis sample included some schools that were eligible to participate in the bonus program but did not ultimately participate. Therefore, the study estimated the effect of being eligible to participate in the program, regardless of actual participation.

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 a 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?

The study found that the New York City Schoolwide Performance Bonus Program had no discernible impact on school Progress Report scores.

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: Because this study examined school-level outcomes, the reported effect sizes are not comparable to effect sizes calculated for student-level analyses.

Appendix A: Study details

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J., Kalra, N., DiMartino, C., Peng, A. (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses. Final evaluation report*. Santa Monica, CA: RAND Corporation.

Setting

The study was conducted in New York City public schools.

Study sample

Four hundred and twenty-seven schools were initially eligible for the program based on low academic performance and demographic characteristics. Of these eligible schools, 25 were removed from the study before random assignment; the report authors were unable to explain the reason for this removal. The remaining 402 schools were then randomly assigned to either an intervention or a comparison group (234 intervention and 168 comparison). For reasons that are not clear, the report uses 399 schools (232 intervention and 167 comparison) as the baseline sample for the analyses. Of these 399 schools, 10 were missing data in Year 1, resulting in an analysis sample of 389 schools. In addition, four schools were dropped from the study in Year 2 because they were no longer operating; no schools dropped out of the study in Year 3.

To participate in the bonus program, 55% of an intervention school's full-time union teachers had to vote in favor of participation. However, because of the "intent-to-treat" study design, all schools randomly assigned to the intervention group were included in the analysis, regardless of whether they participated in the bonus program. For all three years of the study, the WWC did not find attrition to be severe enough to question the findings pertaining to program effects.

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. At the end of the school year, the NYCDOE 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 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

Outcomes were measured using five scores from the NYCDOE's Progress Reports: Environment, Performance, Progress, Additional Credit, and Overall, which is a weighted average of environment (15%), performance (25%), and progress (60%), plus additional credit. For further details, 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 eligible for review by the WWC by receiving significant media attention.

Appendix B: Outcome measures for the academic achievement domain

Academic achievement	
Environment score on New York City's Department of Education (NYCDOE's) Progress Report	The Environment score is based on factors such as student attendance and the results of NYCDOE-issued teacher, parent, and student surveys (for middle and high schools only) that measure perceptions about academic expectations, communication, engagement, safety, and respect at the school.
Performance score on NYCDOE's Progress Report	For elementary and middle schools, the Performance score is based on students' annual scores on the New York state tests in English language arts and mathematics. For high schools, the score is based on graduation rates.
Progress score on NYCDOE's Progress Report	For elementary and middle schools, the Progress score is based on average school improvement on the state test from the previous year. For high schools, it is based on credit accumulation and completion of weighted pass rates for the Regents Examinations. Additional credit is awarded for exemplary progress with high-needs populations.
Additional Credit score on NYCDOE's Progress Report	The NYCDOE's Progress Report awards additional credit to schools that show exemplary progress with high-needs populations.
Overall score on NYCDOE's Progress Report	The Overall score is a a weighted average of environment (15%), performance (25%), and progress (60%), plus additional credit.

Appendix C.1: Study findings for Year 1

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	
Academic achievement (school-level)								
Environment	Elementary, Middle, and High schools	389 schools	nr	nr	–0.06	–0.02	–1	> 0.05
Performance	Elementary, Middle, and High schools	389 schools	nr	nr	0.15	0.04	+1	> 0.05
Progress	Elementary, Middle, and High schools	389 schools	nr	nr	–0.28	–0.03	–1	> 0.05
Additional Credit	Elementary, Middle, and High schools	389 schools	nr	nr	–0.20	–0.07	–3	> 0.05
Overall	Elementary, Middle, and High schools	389 schools	nr	nr	–0.39	–0.02	–1	> 0.05
Domain average for academic achievement in Year 1						**–0.02**	**–1**	**Not 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 an average school's outcome that can be expected if the school is given the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in an average school's percentile rank that can be expected if the school is given the intervention. The WWC-computed average effect size is a simple average rounded to two decimal places; the average improvement index is calculated from the average effect size. Because this study examined school-level outcomes, these effect sizes are not comparable to those calculated for student-level outcomes. The statistical significance of the study's domain average was determined by the WWC; the study did not show any discernible effects of the program on school-level academic achievement in Year 1 of program implementation because none of the estimated effects were statistically significant. nr = not reported.

Study Notes: No corrections for clustering or multiple comparisons were needed. The *p*-values presented here were reported in the original study.

Appendix C.2: Study findings for Year 2

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	
Academic achievement (school-level)								
Environment	Elementary, Middle, and High schools	371 schools	nr	nr	0.14	0.06	+2	> 0.05
Performance	Elementary, Middle, and High schools	371 schools	nr	nr	0.00	0.00	0	> 0.05
Progress	Elementary, Middle, and High schools	371 schools	nr	nr	0.01	0.00	0	> 0.05
Additional Credit	Elementary, Middle, and High schools	371 schools	nr	nr	–0.03	–0.01	0	> 0.05
Overall	Elementary, Middle, and High schools	371 schools	nr	nr	0.11	0.01	0	> 0.05
Domain average for academic achievement in Year 2						**0.01**	**0**	**Not 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 an average school's outcome that can be expected if the school is given the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in an average school's percentile rank that can be expected if the school is given the intervention. The WWC-computed average effect size is a simple average rounded to two decimal places; the average improvement index is calculated from the average effect size. Because this study examined school-level outcomes, these effect sizes are not comparable to those calculated for student-level outcomes. The statistical significance of the study's domain average was determined by the WWC; the study did not show any discernible effects of the program on school-level academic achievement in Year 2 of program implementation because none of the estimated effects were statistically significant. nr = not reported.

Study Notes: No corrections for clustering or multiple comparisons were needed. The *p*-values presented here were reported in the original study.

Appendix C.3: Study findings for Year 3

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	
Academic achievement (school-level)								
Environment	Elementary, Middle, and High schools	371 schools	nr	nr	–0.12	–0.04	–2	> 0.05
Performance	Elementary, Middle, and High schools	371 schools	nr	nr	–0.31	–0.06	–2	> 0.05
Progress	Elementary, Middle, and High schools	371 schools	nr	nr	–0.60	–0.05	–2	> 0.05
Additional Credit	Elementary, Middle, and High schools	371 schools	nr	nr	–0.29	–0.11	–4	> 0.05
Overall	Elementary, Middle, and High schools	371 schools	nr	nr	–1.32	–0.07	–3	> 0.05
Domain average for academic achievement in Year 3						**–0.07**	**–3**	**Not 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 an average school's outcome that can be expected if the school is given the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in an average school's percentile rank that can be expected if the school is given the intervention. The WWC-computed average effect size is a simple average rounded to two decimal places; the average improvement index is calculated from the average effect size. Because this study examined school-level outcomes, these effect sizes are not comparable to those calculated for student-level outcomes. The statistical significance of the study's domain average was determined by the WWC; the study did not show any discernible effects of the program on school-level academic achievement in Year 3 of program implementation because none of the estimated effects were statistically significant. nr = not reported.

Study Notes: No corrections for clustering or multiple comparisons were needed. The *p*-values presented here were reported in the original study.

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 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] The study also examined academic achievement outcomes measured at the student level. However, the report did not contain enough information to determine a study rating for that portion of the study.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2013, September). *WWC review of the report: A big apple for educators: New York City's experiment with schoolwide performance bonuses. Final evaluation report.* 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.