Abstract Title Page
Not included in page count.

Title: Evaluating Phase II of a New York City-Wide STEM Initiative using Propensity Score Methods: A Replication Study.

Authors and Affiliations:

Ally S. Thomas, Graduate School & University Center, City University of New York, astevens@gc.cuny.edu

Sarah M. Bonner, Hunter College, City University of New York, sbonner@hunter.cuny.edu

Howard T. Everson, Graduate School & University Center, City University of New York, HEverson@gc.cuny.edu

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Recently, we have been exploring the use of propensity score methods for developing evidence of program impact. Specifically, we have been developing evidence (after one year of implementation) of the effects of the Math Science Partnership in New York City (*MSPinNYC2)* on high school students' achievement—both in terms of course grades and scores on end-of-course tests in two key Science, Technology, Engineering and Mathematics (STEM) disciplines: Integrated Algebra and Living Environment. Using an evidence-based approach which relies on *propensity score matching* (Rosenbaum & Rubin, 1983; 1984; Rubin, 2006), we asked if the program in its early stages is making a difference in students' academic achievement and college readiness.

The *MSPinNYC2* program restructures early high school STEM courses to include 6-8 Teaching Assistant Scholars (TAS) who, along with the teachers, facilitate in-classroom group work on a daily basis. Early pilot studies suggested the model of peer-enabled restructured classrooms (PERC) increases student achievement and narrows the achievement gap in high school STEM courses. In its inaugural year the *MSPinNYC2* project recruited over 700 students (n = 711) in four New York City public high schools to participate in PERC classes. The students served by the project are not the academically elite (e.g., only about 20-25% were proficient in math and/or English language arts at the end of 8th grade). In the first year (2011-12) the PERC courses were taught by eleven different high school teachers—all trained in the PERC pedagogical model by the Program's staff. Each course served between 25-30 students who were tutored in-class by 6 or 7 TAS.

Propensity score matching (PSM) methods were used to evaluate the efficacy of the *MSPinNYC2* for PERC students after this first year of implementation. PSM had its origins nearly three decades ago in biomedical research for reducing estimation bias when comparing non-equivalent groups, and for drawing causal inferences in observational study designs (i.e., studies where random assignment to treatment is not possible).

Results from the first year suggest the *MSPinNYC2* was not effective in raising academic achievement for PERC students in the 9th grade mathematics course (Integrated Algebra), but was effective for PERC students in the 9th grade biology (Living Environment) course. Furthermore the study provided evidence that PSM is valuable and effective in monitoring the efficacy of a large multi-site instructional intervention.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

Because the *MSPinNYC2* program is a multi-year, multi-site STEM intervention, with new cohorts of students entering the program each year, it is essential to develop a sound, rigorous method for evaluating the effectiveness and the scalability of the intervention—one that can be replicated each year. We are keenly interested in models that allow us to replicate our results

from year-to-year with each new cohort of students. Thus, the purpose of our proposed study is to explore the use of a propensity score model that uses a genetic matching algorithm to evaluate the replicability of our limited first-year findings, using as participants a new cohort of students who participated in the second year of the program. Specifically, we examine (1) the effectiveness of using propensity score methods to evaluate a math/science educational intervention and (2) the replicability of the findings on a new cohort of students.

Setting:
Description of the research location.

The *MSPinNYC2* program is a city-wide STEM initiative that has been implemented in six New York City public high schools. The collaborative research and development team studying the implementation and efficacy of the intervention come from Hunter College, CUNY and the Center for Advanced Study in Education at the Graduate School, CUNY. Data collection and database development and analyses are conducted in conjunction with New York City's Department of Education (DOE).

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.

Participants in this study will include the cohort of students who participated in first and second years of the *MSPinNYC2* program. This includes roughly 2000 PERC students from six New York City public high schools. The majority of students served are Black (38%) and Hispanic (47%), and many struggle as the result of poor academic preparation. Again, the students served by the project are not the academically elite (e.g., only about 20-25% were proficient in math and/or English language arts). For purposes of creating a matched sample, extensive student background and achievement data will be provided on a sample of roughly 10,000 9th grade students randomly selected from the larger population of all NYC high schools.

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.

The engine of the *MSPinNYC2* program is the Peer Enabled Restructured Classroom (PERC**).** This pedagogical approach appears to be a powerful innovation in mathematics and science instruction—one that supports teachers, improves student performance, creates young leaders, and engages students deeply in learning. Ninth graders enrolled in math (Integrated Algebra) and science (Living Environment) work together daily in small groups (four to five students) solving problems, completing assignments and mastering required curriculum with the guidance of trained group leaders, the TAS.

Research Design:
Description of the research design.

The current study uses an observational design and is a replication of the initial, preliminary propensity score analysis we conducted at the end of the first year of the program. Below we describe the variables we used, and will continue to use, during the replication phase of our program evaluation efforts.

Dependent variables. Because we propose a replication design, we again will use the same dependent variables as in our earlier, Phase I work. The outcome variables will include students' scores on the Integrated Algebra and/or the Living Environment Regents exams. These standardized, end-of-course exams are scored on a 0-100 scale. In addition to using a continuous score scale, we will also include binary (pass/fail) variables, operationalized in the state of New York as a score of 65 or higher on each exam, as well as binary college readiness indicators. As college readiness indicators, we use a score of 80 or higher on the Integrated Algebra Regents exam, based on City University of New York expectations, and a score of 75 or higher on the Living Environment Regents exam. A more complete description of the New York State Regents exams can be found at www.nysedregents.org.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

*Data Collection***.** The data used for these analyses have been collected under a cooperative agreement with the New York City DOE and the City University of New York. All data have been de-identified. Students' background variables and prior academic achievement indicators include gender, age, grade-level, ethnicity, and 8th grade NYC state achievement test scores in English language arts and mathematics.

*Calculating propensity scores***.** The first step in our analyses is to calculate the students' propensity score, i.e., the predicted probabilities for each student that they received the treatment. It is calculated using logistic regression where the dependent variable is the treatment and the predictors are all of the presumably related covariates. For purposes of replication, we will use the same covariates as used in Phase I: age, race/ethnicity, gender, English as a first language indicator, a free or reduced lunch indicator, attendance, 8th grade math scores and 8th grade ELA scores.

Matching. Once we obtain the propensity scores we will then use those scores to match treated and non-treated students. As noted earlier, we will use a genetic matching method—one used iteratively to check and improve the covariate balance resulting in an optimal balance of all the relevant covariates (Diamon & Sekhon, in press).

*Regression-adjusted matched estimates***.** In order to estimate and replicate the treatment effect, a regression of the outcomes on the treatment indicator and confounding covariates will be used. The regression models will be adjusted using weights to force the sample to represent the treated group of students.

Findings / Results:
Description of the main findings with specific details.

Phase II of the replication study is currently underway. Below we describe the results from Phase I which was conducted at the close of the first full-year of implementation of the *MSPinNYC2*.

Integrated Algebra (IA). Three different analyses were conducted to test three hypotheses: (1) students in the PERC IA course would score higher on average on the end-of-course exam than non-PERC IA students. Results from the Phase I regression adjusted matched estimate suggest no significant treatment effect, β = -0.173, p = .771. (2) PERC IA students are more likely to pass the IA exam than non-PERC students. Again, results from the logistic regression adjusted matched estimate suggested no significant treatment effect, β = 0.0169, p =

.906. And (3) PERC IA students would be more likely to score 80 or higher (the college ready benchmark) on the IA exam than non-PERC students. Results from the logistic regression adjusted matched estimate suggested no significant treatment effect, $\beta = 0.0359$, $p = .889$.

Living Environment (LE, Biology). Similar to the IA approach, three analyses were conducted to test the same three hypotheses for the PERC students in the Living Environment (LE) course. Results from the regression adjusted matched estimate suggest a significant treatment effect, $\beta = 1.946$, $p = .008$, indicating that PERC students in the LE course scored, on average, 1.95 points higher on the LE exam than non-PERC students. Second, the PERC LE students were more likely to pass the LE Regents exam than non-PERC students, $\beta = 0.502$, $p = .0070$. Estimating the magnitude of this effect, PERC LE students were 1.65 times more likely to score a 65 (passing score) or higher on the LE exam than non-PERC students. And third, the PERC LE students were expected to be more likely to score a 75 (the college ready benchmark) or above on the LE Regents exam than non-PERC students. Results of the logistic regression adjusted matched estimate indicated no significant treatment effect, $\beta = 0.255$, $p = .199$, for the third analysis.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

We expect the estimates from the Phase I propensity score analyses will be replicated in Phase II. Because the Phase I model results were not available to feed back to program leadership until near the end of the second year of implementation, we expect that changes in implementation of the PERC model in the second year were minor. Thus we predict our estimates from Phase I will be replicated in Phase II. This stability in program implementation provides an excellent opportunity to evaluate consistencies among different estimates using the same methodology. We propose to use two criteria to determine whether a given Phase II impact estimate replicates the initial Phase I benchmark. The first criterion we will consider is whether the conclusion that would be drawn from the Phase II impact estimate is the same as for Phase I across both the algebra and biology courses. Specifically, we propose to examine whether the basic magnitude and sign of the estimates (and the estimated odds) are comparable and whether the statistical significance (or insignificance) is the same as in Phase I. The second criterion is whether the Phase II non-experimental estimates are statistically different from the Phase I benchmarks. Our results will be discussed in terms of the implications for using propensity score methods to evaluate relatively large-scale STEM interventions, and to provide robust evidence of replicability and going to scale with such intervention

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Diamond, A. & Sekhon, J. S. (in press). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* (forthcoming).

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(*1*), 41-55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of American Statistical Association*, 79(*387*), 516-524.

Rubin, D. R. (2006) *Matched sampling for causal effects.* NewYork, New York: Cambridge University Press.