Abstract Title Page

Title: Assessing Early Impacts of School-of-One: Evidence from the Three School-Wide Pilots

Author(s):

James J. Kemple
James.Kemple@nyu.edu
Research Alliance for New York City Schools

Micha D. Segeritz
Michael.Segeritz@nyu.edu
Research Alliance for New York City Schools

Rachel Cole
Rachel.Cole@nyu.edu
Research Alliance for New York City Schools

Section:

Mathematics and Science Education in the Secondary Grades

Abstract Body

Background / Context:

US students are not performing strongly in math. Only 39% of fourth graders, 34% of eighth graders, and 26% of twelfth graders performed at or above proficiency on the National Assessment of Educational Progress (NCES, 2009; NCES, 2010). Progress on successive NAEP tests has been meager and racial achievement gaps persist. Internationally, the 2009 average score of US students on the Program for International Student Assessment (PISA) was lower than the average score for OECD countries, despite small gains since the 2006 assessment (OECD, 2010).

In the development of math skills over the course of primary and secondary education, the middle years are a key time when students' math performance begins to lag (Lee and Fish, 2010). Increasingly educators and researchers are seeking out instructional methods that allow teachers to meet students' individual needs (Davis, 2011). Students come to classroom with varying levels of prior understanding, needing instruction in different skills, and with diverse interests and preferred learning styles. So1 is a new, individualized, technology-rich math program being implemented in three New York City middle schools. The program offers a high level of customization for each student, both in the content and material with which students engage, and in the teaching and learning modalities that are used to enhance students' mastery of the material. Student, teacher, and parent surveys as well as ongoing assessment information inform how the program is tailored for each student. So1 is the recipient of a three-year, five million dollar Investing in Innovation (I3) development grant from the federal Department of Education, and was named one of the top fifty inventions of 2009 by Time magazine.

The proposed paper evaluates the effectiveness of So1 in improving math test scores in its first year of complete implementation. While the study adds to previous research on So1 conducted by the Education Development Center for Children and Technology (CCT) and the New York City Department of Education (NYCDOE), our study is the first independent evaluation of this new, expanding program. Our study's preliminary assessment of impact will contribute to program development for future I3 development as So1 continues to evolve.

Purpose / Objective / Research Question / Focus of Study:

- What is the impact of the initial whole school version of So1 on students' math achievement: 1) as measured by performance on the New York State assessment in mathematics, and 2) as measured by performance on the interim, periodic mathematics assessments provided by Acuity?
- To what extent does the impact of So1 differ across subgroups of students, including those defined by: 1) grade level; 2) prior mathematics achievement; 3) English language learning status; and 4) special education status?

Setting:

New York City Department of Education includes over 1500 schools serving more than 1.1 million students. So1 is being implemented in three NYC middle schools: We shall refer to these schools as Manhattan School, Brooklyn School, and Bronx School. See Table 1 for descriptive statistics on these schools.

Population / Participants / Subjects:
So1 replaced the traditional math instruction and curriculum for sixth, seventh and eighth graders in Manhattan School and Brooklyn, and for sixth graders and some seventh graders in Bronx School. The program serves all students with the exception of newly arrived English Language Learners (ELLs) and students with special education needs requiring small, self-contained classrooms. In total, So1 serves approximately 1,700 students.

Intervention / Program / Practice:
So1 is a program of mass customization of student learning in response to the diverse levels of math proficiency and different preferred learning modalities students bring with them. The program develops a "playlist" of the math skills a particular student needs to work on based on an automated analysis of a variety of assessment data. Data from a number of surveys about the student's interests and learning style form a student profile that will influence the kinds of lessons that are presented to the child. A lesson bank offers lessons to address each skill in a number of different formats: large group, small group, virtual tutoring, collaborative group activities, and educational software. A learning algorithm matches each student's profile to the required instructional content to generate a schedule for all students and teachers every school day. Every day the students are assessed further, and these data feed back into the algorithm to improve it and to add more information to the student's profile. The schedule the algorithm generates for each student each day is flexible, however, allowing teachers to adjust the schedule as necessary for their students.

Before this year's first full implementation of So1, the program went through a number of pilot steps. In summer 2009, So1 provided a four week long summer school program for 80 rising seventh graders. In spring 2010, it was adapted as an after school program for 240 sixth graders during seven weeks. Later that spring So1 became the school-day math instructional program for 200 sixth graders during six weeks.

Research Design:
We estimate the impact of So1 on math achievement using a comparative interrupted times series methodology, a method commonly used to evaluate school-wide programs.[1] The first stage of this method is to construct a baseline model representing the trend in math achievement in each grade at each school. We use data from 2006 through 2010 to construct this baseline trend. Extending this trend one further year gives an estimate of what the scores might have been if So1 had not been implemented. The difference between this predicted score and the actual 2011 score is an estimate of the impact of So1. However, we cannot necessarily attribute this deviation to So1 since other district-wide reforms and policies may have come on line during this period and influenced math achievement independent of So1.

The second stage of the analysis begins with the identification of a set of comparison schools. We use a combination of test-score trajectories from 2006 to 2010[2] and student demographics[3] to

[1] This section draws highly on Howard Bloom's methodological work (1995, 1999, and 2003).

[2] The test scores and trajectories are based on cross-section longitudinal regression analysis. For all eligible schools, we use student-level math test scores between 2006 and 2010 to estimate grade specific intercepts in 2006 and year-to-year linear trends through 2010.

[3] These characteristics are: the percent of students who are English language learners, female, the percent who are Hispanic/Black, or Asian, the percent that have a special education designation (including only students that participated in the math test), the percent of students eligible for free and reduced price lunch, and the school-wide attendance rate.

identify six comparison schools for each So1 school.[4] For potential comparison schools, we construct an index that measures the similarity to each So1 school and chose the six schools with the closest scores to form the comparison group.[5] For these schools, we conduct the same interrupted time-series regression analysis for the comparison schools that is described above in Stage 1 for the So1 schools. Since these schools were not exposed to So1, deviations from the baseline trend would be due to other reforms or initiatives being implement across the district or in selected middle schools like these.

In the final stage of the analysis, we compare the differences estimated in Stage 1 with the differences estimated in Stage 2. Here, we estimate the difference between the test scores predicted by the baseline trend and the actual test scores in 2011 for the So1 and comparison schools simultaneously (combining step 1 and 2 in one model). Thus, this so-called "difference-in-difference" approach contrasts the gains in 2011 for the So1 schools to the gains for the comparison schools. This estimate represents the best indication of the impact So1 has on student math test scores over and above the influence of prior initiatives and trends and simultaneous interventions that may be underway across the district or in schools like those being served by So1.

In addition to the whole population of 6th, 7th and 8th graders in the selected schools, we will also estimate the impact of So1 for different subgroups of students separately. The analytic strategy for these analyses are the same as the strategy described above except that we will focus on discrete subgroups of students defined by prior performance levels and demographic characteristics.

The minimum detectable effect size (MDES) for the full sample of three schools is between .24 and .45 effect sizes. For a single grade, a single school, or the students of a subgroup (pooled across schools) the MDES is likely to range between .50 and .82 effect sizes. While these are moderate to large effect sizes, we nonetheless think it is entirely possible that we will detect an impact. For the pooled sample, scores would need to rise an average of 4 to 7.5 scale score points, which is roughly equivalent to the citywide growth New York City has witnessed over the past two to three years. So1 is a dramatically different mechanism for the delivery of math instruction and may have a large enough impact to be detected.

Data Collection and Analysis:

Our data sources include New York State math test scores from 2006 through 2011. In addition, we will use Acuity math predictive scores (a low-stakes, periodic assessment meant to inform instruction and widely used in schools across the city) from early 2011. We draw demographic information on school composition from the J Form, a publically-available dataset.

Our analysis will estimate a linear baseline comparative interrupted time-series model:

$$Y_{ijt} = INT + \beta_1 * YR_{ijt} + \beta_2 * Post2010_{ijt} + \beta_3 * X_{ijt} + \beta_4 * So1_{ijt}$$
$$+ \beta_5 * So1_{ijt} * YR_{ijt} + \beta_6 * So1_{ijt} * Post2010_{ijt} + \varepsilon_{ijt}$$

[4] Potential comparison schools included the 189 New York City schools with a middle school grade configuration (Grades 6-8) that operated continuously between 2006 and 2011.

[5] Based on the concept of "Euclidian distances" used in many cluster analyses, the similarity index captures the multi-dimensional differences between each So1 school and each potential comparison school based on important background and performance characteristics.

Where:

Y_{ijt} = Test score for student *i* in school *j* in year *t*

YR_{ijy} = Year of observation for student *i* in school *j*, where -3, -2, -1, and 0 correspond to 2006 - 2010, respectively and 1 corresponds to 2011

$Post2010_{ijt}$ = 1 if observation for student *i* in school *j* is from 2011, 0 if observation is from 2006-2010

X_{ijt} = Vector of predictors of individual student characteristics for student *i* in school *j* in year *t*

$So1_{ijt}$ = 1 if school *j* is a So1 school, 0 otherwise

$So1_{ijt} * YR_{ijt}$ = Year of observation for So1 school *j*, where -3, -2, -1, and 0 correspond to 2006 - 2010, respectively and 1 corresponds to 2011

$So1_{ijt} * Post2010_{ijt}$ = 1 if school *j* is a So1 school and the year is 2011, 0 otherwise

ε_{ijt} = random error for student *i* in school *j* in year *t*

The parameter estimate β_6 represents the estimated difference in test scores between So1 schools and non-So1 schools in 2011 period accounting for differences in trends between the groups during the period 2006-2010 and accounting for differences in the demographic characteristics in the two groups of schools. We will run this analysis three times, once for each So1 school and its comparison schools, and then pool these results to form one estimate of the program impact.

Findings / Results/Conclusions:
Findings and conclusions with regard to our research questions are forthcoming. Figure 1 shows two findings from our initial analysis that suggest our research design is particularly strong. First, the baseline trends for both the So1 schools and the comparison schools are highly predictive of the observed data, and thus will serve as a strong predictor of the 2011 scores. Second, the baseline trends for the So1 schools and the comparison schools show a high degree of similarity both in the levels of student performance and in the year-to-year growth. Because of this similarity, we may have a high degree of confidence that subsequent differences that may emerge between the schools in 2011 are likely to be due to one school being exposed to So1 and the comparison schools not being exposed. From these two findings, we conclude that our research design is a good choice for these data. Table 1 gives the precise numbers displayed in Figure 1, as well as some demographic data about the two groups.

One potential limitation of our study stems from So1's recruitment process. The schools in which it was implemented were purposefully identified as being good candidates for successful implementation, so we cannot rule out the possibility that any effect we find could be due to selection bias. A final limitation is our limited power, which yields a MDES of .24 and .45 effect sizes.

Appendices

Not included in page count.

Appendix A. References

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19*(5), 547-556.

Bloom, H. S. (1999). *Estimation program impacts on student achievement using "short" interrupted time series*. New York: MDRC. Retrieved September 2, 2010, from http://www.mdrc.org/publications/82/full.pdf

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms. *Evaluation Review, 27*(1), 3-49.

Davis, M. R. (2011). “Moving Beyond One-Size-Fits-All.” *Education Week*. Bethesda: Mar 17, 2011. Vol. 30, Iss. 25; pg. 10.

Education Development Center for Children and Technology (2009). *Evaluation of the School of One Summer Pilot: An Experiment in Individualized Instruction.* New York, New York.

Lee, J., and Fish, R. M. (2010). International and Interstate Gaps in Value-Added Math Achievement: Multilevel Instrumental Variable Analysis of Age Effect and Grade Effect. *American Journal of Education*, Vol. 117, No. 1 (November 2010), pp. 109-137

NCES (2010). *The Nation’s Report Card: Grade 12 Reading and Mathematics 2009 National and Pilot State Results.*

NCES (2009).*The Nation's Report Card: Mathematics 2009*.

New York City Department of Education (2010). *School of One Evaluation – 2010 Spring Afterschool and Short-Term In-School Pilot Programs*. New York, New York.

OECD (2010). *PISA 2009 at a Glance,* OECD Publishing. http://dx.doi.org/10.1787/9789264095298-en

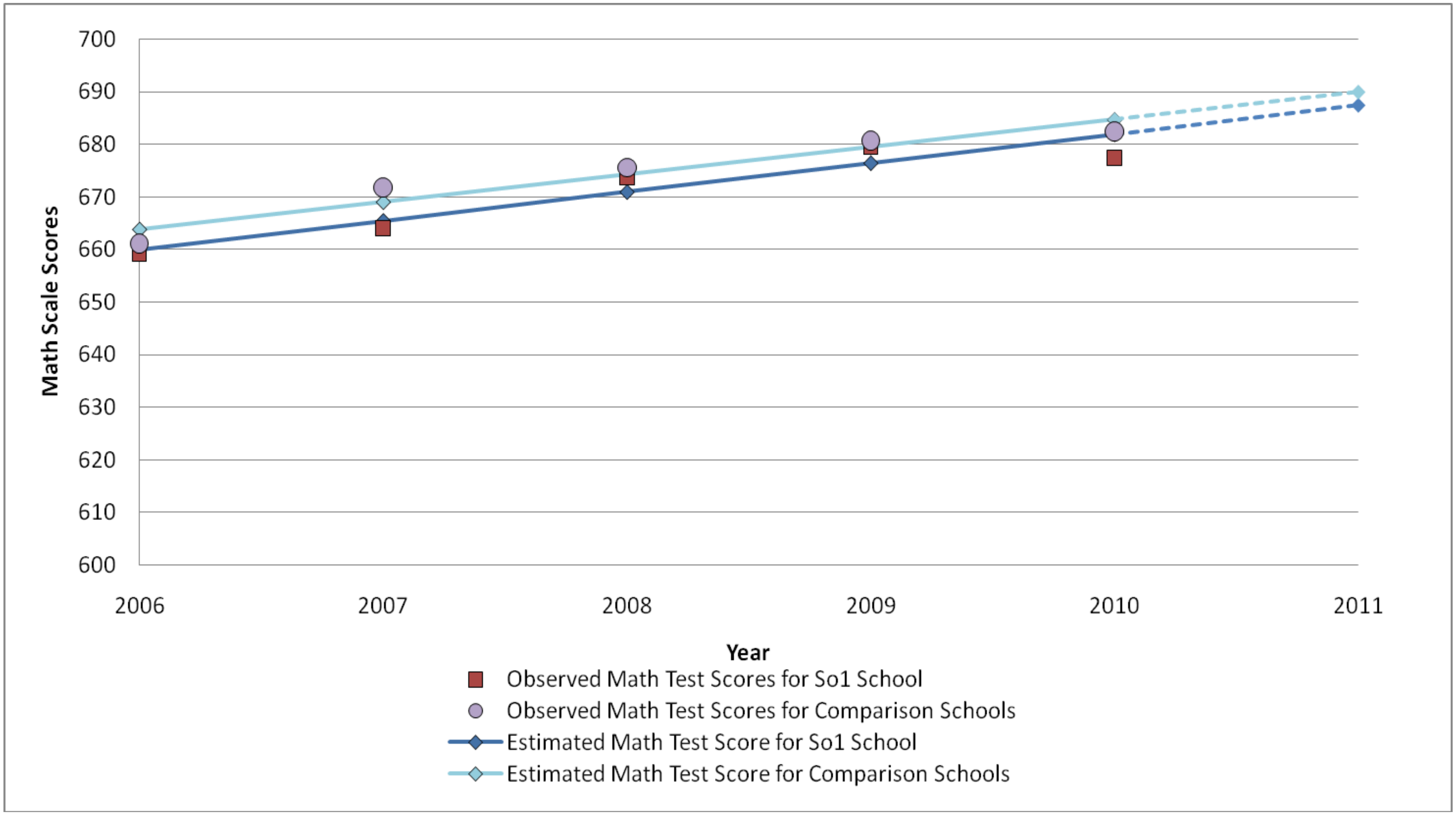
Appendix B. Tables and Figures

Table 1
Characteristics of So1 Schools and their Comparison Schools
Averages for 2006-2010

Characteristic	**So1 Schools**	**Comparison Schools**
Female (%)	44.2%	48.6%
Race/ethnicity (%)		
White	10.0%	11.7%
Black	17.0%	18.3%
Hispanic	32.6%	43.6%
Asian	40.0%	25.7%
English language learners (%)	26.8%	16.1%
Special education (%)	7.2%	6.0%
Poverty[6] (%)	75.2	68.4
Peer index	2.7	2.7
Observed math scores in 2010	669.5	672.5
Estimated math scores in 2010	671.6	676.4
Estimated yearly change in math scores between 2006 and 2010	7.5	7.8

[6] In 2009.

Figure 1:
Baseline Math Test Score Trend and Projection
Grade 6, Brooklyn School



This chart demonstrates how well-suited the comparative interrupted time series methodology is to the present question. First, the predicted baseline models—both for the So1 schools in and for the comparison schools—are quite close to the actual observed scores. Second, the models for the So1 schools and the comparison schools are quite close to one another. Taken together, these two findings support the assumptions on which the comparative interrupted time series analysis is based.