Rolling Out and Scaling Up:
What Happens When a New Program Is No Longer New

Meryle Weinstein[1,2] and Emilyn Ruble Whitesell[1,3]

[1]Institute for Education and Social Policy, New York University
[2]Steinhardt School of Culture, Education, and Human Development, New York University
[3]Wagner Graduate School of Public Service, New York University

Prepared for the XXIV Meeting of the Economics of Education Association
(Asociación de Economía de la Educación, AEDE)
Madrid, Spain

June 2015

I. Introduction

A major focus of U.S. education policy conversations in recent years has been the role of innovation. With the goal of transforming the education landscape, many policymakers and practitioners are intent on identifying new programs and pedagogical approaches, from developing new education governance structures to leveraging technology. Despite the common belief that the education sector has stagnated, it features considerable innovation, both in absolute terms and relative to other industries (Fisher & Vincent-Lancrin, 2014).

Funding streams such as the federal government's Investing in Innovation (i3) grants support educational innovation, with an emphasis on scaling up programs that are shown to be effective (U.S. Department of Education, 2015). If programs do not show positive effects within two or three years of implementation, they are often discarded. This is problematic for several reasons. First, programs typically take at least three years to become effective (Fullan & Stiegelbauer, 1991), and many potentially effective initiatives may be abandoned before they have had sufficient time to "work." As start-up costs can be substantial, this can be an inefficient use of resources. Second, the continual adoption and dismissal of education programs contributes to considerable flux within school systems and individual schools (Stevens, 2004). It may be difficult to get buy-in from staff and students for new approaches when programs are constantly coming and going.

Additionally, we know very little about how successful programs have been able to scale up, other than major national programs such as Teach For America or Success for All. Local and regional programs that operate on a large scale throughout a school district, while key providers of educational resources for many students, are under-studied. Barriers to studying these types of programs include data limitations, sufficient scale or sample size to rigorously estimate effects,

and – of course – funding. In this study, we have the unique opportunity to explore a large-scale program that has been operating in New York City (NYC) for more than a decade. Urban Advantage (UA) is a formal-informal partnership in NYC that incorporates the resources of NYC's informal science education institutions (ISEIs) and the New York City public school system to improve instruction in middle school science. UA provides intensive professional development for participating teachers, materials for science classrooms, and free access to ISEIs for class trips and independent visits. This program is designed specifically for NYC and is more closely integrated with the city's science curriculum than typical formal-informal programs that focus on general science enrichment.

Now in its eleventh year of operation, UA has grown and become embedded in NYC's approach to science instruction. In the 2013-14 (hereafter 2014) academic year, more than 30% of all NYC middle schools were actively participating in UA and roughly 50% of middle schools had ever participated in UA. This presents a unique opportunity to evaluate a program that started small, has grown to scale over a period of time, and has become a key component of the New York City Department of Education's (NYCDOE) middle school science professional development. Importantly, by linking program data to administrative data from the NYCDOE and the New York State Education Department (NYSED), we are able to explore changes in program effects as the program has grown and persisted in NYC.

In a previous study of the impact of UA on students' science outcomes (Weinstein, Whitesell, and Schwartz, 2014), we found that attending a UA school increases students' performance on the eighth-grade science exam by approximately 0.05 standard deviations, with larger effects for black students, students in special education programs, and boys. We found small positive effects on the likelihood that a student will take a science Regents exam in eighth

or ninth grade but no consistent effect of UA on a student's probability of being proficient on a science Regents exam. Finally, we did not find any systematic effect of attending a UA school in eighth grade on a student's likelihood of attending a science, technology, engineering, and math (STEM) high school. This previous study used data from 2004-2010 (six years of UA implementation) and did little to capitalize on variation in schools' program implementation over time. For example, the previous analyses estimated separate effects for a school's first year in UA and subsequent "post" years, but it did not distinguish among different post years or explore repeated exposure to UA at the student level.

A key difficulty of measuring program effects as programs grow to scale is that treatment and comparison groups can become muddled, and in our case schools join and exit UA over time. This means that in any given year, there will be a group of schools actively participating, a group of schools that previously participated but is not currently active, and a group of schools that has never participated; furthermore, with a 10-year-old program, there is substantial variation in the number of years a school has been participating in UA. To use a conservative definition of the treatment, our main models estimate the impact of attending a school that has ever had a UA teacher; for several reasons addressed in later sections of this paper, this biases our estimates toward zero and means we are likely underestimating the true program effect. We also explore heterogeneity in the impact of UA based on the number of years since the school joined.

Another issue to consider with a long-term program is that as schools enter and exit the program, the characteristics of UA schools and students may change considerably, and UA may be more effective for some groups of schools and students. To address this, we estimate our models on subgroups of schools and students. Specifically, we explore heterogeneity in the

impact of UA by schools' prior performance in science and by the percentage of science teachers who are participating in UA. We also estimate traditional subgroup models to determine whether UA is particularly effective for different groups of students.

The rest of this paper is structured as follows. In section II, we provide context about the Urban Advantage program and relevant prior literature. In section III, we describe our data and models, and we report results in section IV. We discuss results and conclude in section V.

II. Context

A. The Urban Advantage Program

Urban Advantage is a formal-informal partnership that unites eight ISEIs (the American Museum of Natural History, Brooklyn Botanic Garden, New York Botanical Garden, New York Hall of Science, Queens Botanical Garden, Staten Island Zoo, and the Bronx Zoo, and the New York Aquarium) with middle schools in NYC to improve inquiry-based science education and ultimately middle school science outcomes. Figure 1 shows the UA logic model, which articulates the activities (or inputs) UA provides (e.g., professional development, access to ISEIs), its proximal outcomes (e.g., greater use of inquiry-based education practices, higher science achievement), and its ultimate goals (e.g., improve students' high school outcomes and college readiness). See Weinstein et al. (2014) for a more detailed description of the Urban Advantage program.

It is important to note that some of the UA inputs are easily retractable once teachers are no longer participating in UA, while others can be thought of as creating more permanent changes in school resources. This will be important in understanding program effects that occur only when teachers are actively participating in UA versus more lasting impacts. For example,

UA provides participating teachers and schools with funds each year to purchase science materials for their classrooms and vouchers to ISEIs; these resources are only available to active teachers and arguably only have contemporaneous effects. On the other hand, schools get an initial investment in their first year to buy equipment which will remain at the school for their use. These science kits include a set of materials geared toward inquiry-based activities and the eighth-grade science exit projects (long-term science investigations all eighth-graders complete before leaving middle school). Schools keep these instructional materials, even if they are no longer participating in UA.

Other UA inputs, however, can be considered investments in the school's human and material capital and may have longer-lasting effects. For example, participating teachers receive intensive professional development that may improve their capacity with inquiry-based teaching approaches in the long term. Teachers new to the program receive 48 hours of professional development across three different cycles in their first year, and teachers continuing in the program receive ten hours of differentiated professional development each year. In these sessions, teachers learn to use UA tools (such as the Designing Scientific Explanations Tool and the Investigation Design Diagram), to implement inquiry-based approaches in their classrooms, and to integrate trips to ISEIs into their curricula. Because it is impossible to "un-treat" teachers who have received UA professional development, as the training may affect their pedagogical approach long term. The situation is further complicated by teacher mobility in New York City. According to UA staff, approximately 1600 teachers have received professional development through UA. Not including retired teachers, there is the possibility that a large number of teachers who received UA professional development are now teaching at non-UA schools.

Without knowing where these teachers are teaching, we cannot say with certainty that "never UA" schools have not had any exposure to UA teaching practices.

Urban Advantage has also changed over time, both in terms of the program itself and the schools it serves. For example, program resources and professional development have evolved since the program began. The balance of teachers participating in UA has shifted from those new to UA to those continuing in the program, and UA has responded by offering more targeted professional development options for continuing teachers. Additionally, UA has provided more opportunities for continuing teachers to take leadership roles, both as "lead teachers," who facilitate certain UA professional development events, and through informal leadership in their schools.

UA has also adjusted its expansion strategy to target schools with high potential for impact. Teachers and schools apply to participate in UA, and over time UA has developed a more rigorous protocol for entry. In the early years, teachers and schools largely self-selected into UA based on interest. In later years, UA established a more defined protocol for accepting schools into the program, due in part to funding constraints. Based on the assumption that UA is most effective in schools with a greater concentration of UA teachers, program staff opted to expand within already-participating schools rather than to increase the number of schools participating in UA. Similarly, while only eighth-grade teachers were invited to participate in UA's first year (2005), UA soon expanded to include seventh-grade (2006) and sixth-grade teachers (2010). Inviting teachers in grades six through eight to participate means that middle school entire science departments can participate in UA. In some schools, all middle school science teachers are in Urban Advantage and students have UA teachers for three years in a row.

Since its inception in 2004-05, Urban Advantage has expanded in New York City and become a feature of the educational landscape, embedded in the district's approach to middle school science education. As shown in Figure 2, the number of schools participating in UA has grown over time. In 2014, 147 schools participated in UA, representing 30% of all NYC public middle schools. Throughout the past decade, schools have moved in and out of program participation, meaning that the number of schools actively participating in UA in any given year understates the number of schools that have been reached by UA. Figure 2 also includes the number of schools that have *ever* participated in UA, by year. By 2014, there were 90 former UA schools, and the 237 schools that had ever participated in UA represented approximately half (48%) of all NYC middle schools.

As shown in Table 1, which provides average characteristics of UA and non-UA schools in 2005 and 2014, characteristics of participating schools have changed over time. Comparing the 26 UA schools in 2005 to the 147 UA schools in 2014, we see key differences in terms of student and school characteristics. Compared to UA schools in 2005, participating schools in 2014 had lower shares of black (32.7% vs. 38.5%) and Hispanic students (42.2% vs. 50.0%), smaller shares of LEP students (11.6% vs. 15.4%), and lower average math and reading proficiency rates (23.0% and 26.8% vs. 41.7% and 34.8%, respectively).[1] UA schools in 2014 had greater shares of Asian (17.0% vs. 7.7%) and white students (8.2% vs. 3.9%) and greater shares of students in special education (16.3% vs. 11.5%) than UA schools in 2005.

Overall, schools participating in UA are similar to non-UA schools, though in any given year there are small differences in characteristics. The one consistent difference in school characteristics is that UA schools tend to be larger than non-UA schools. For example, in 2005

[1] Note that passing rates on standardized tests fell dramatically across the city in 2013, due to changes in the content and grading of standardized exams. This is reflected in the table, as math and reading test scores for non-UA schools also dramatically fell from 2005-2014.

UA schools had an average of 1,178 students, versus 828 students in non-UA schools; in 2014, UA schools had an average of 730 students, compared to 591 students in non-UA schools. There are no other statistically significant differences between UA and non-UA schools in 2005, and in 2014, the only other statistically significant difference is that UA schools have higher shares of Asian students (17.0% vs. 9.0%).

As shown in Figure 3, the schools entering UA in more recent years have relatively low histories of science achievement. These differences in the characteristics of schools joining UA in recent years indicate that schools new to the program may have different needs than those joining in prior years. Note, however, that in some years very few schools are new to UA (e.g., 3 new UA schools in 2011), and so differences in the characteristics of schools joining UA are likely not driving differences in the characteristics of UA schools overall.

Figure 3, Panel A, shows the number of new UA schools in each year, the average science proficiency rates for new schools in the year before they joined UA, and the average proficiency rate for all middle schools in the prior year. From 2005 to 2010, schools new to UA had prior science performance that was similar to the overall average. In 2011, however, the three schools new to UA had significantly lower lagged proficiency rates (19% vs. 54%). The schools joining UA in 2012 had similar prior performance to middle schools overall (53% vs. 52%), but again in 2013 and 2014, schools new to UA had somewhat lower previous proficiency (45% vs. 54% in 2013, 50% vs. 56% in 2014).

Panel B shows similar information, this time measuring schools' prior science performance with average z-scores instead of proficiency rates. Because this graph uses school-level data (mean z-scores within schools), the performance of middle schools overall is approximately but not exactly 0.00 in all years. From 2005 to 2010, schools new to UA had

lagged performance similar to and in several cases higher than the citywide average. In 2011, however, the three schools new to UA had significantly lower average z-scores in the prior year, compared to middle schools overall (-1.05 vs. -0.09 SDs). For schools joining from 2012-2014 average z-scores in the year before joining UA are also slightly lower than middle schools overall (-0.15 vs. -0.009 in 2012, -0.24 vs. -0.02 in 2013, and -0.15 vs. -0.01 in 2014).

B. Relevant Literature

This study focuses on a large-scale program that has become embedded in the NYC public schools. As we seek to understand how the program effects have changed over time, it is useful to consider what it means for a program to become institutionalized. Growing to scale and institutionalization may occur simultaneously but are distinct concepts; specifically, a program is said to be institutionalized when it is well-integrated into the overall organization and is viable in the long term (Steckler & Goodman, 1989). In early work on institutionalization of programs in higher education, Clark (1968) outlines four different models of institutionalization. Each of these models describes how innovations grow in complexity, become systematized, and are strengthened; these models are the organic growth model, the differentiation model, the diffusion model, and the combined-process model. Central to both the diffusion and combined-process model are the stages of evaluation, trial, and adoption. Applied to the Urban Advantage program, these models suggest that after initial evaluation and a relatively small-scale implementation (i.e., the initial years of UA) the program is adopted on a larger scale and a more permanent basis. Even after adoption, however, the program will be continually re-evaluated (Clark, 1968).

Using diffusion theory, one study of how a health program was implemented and maintained in schools suggests that institutionalization is more likely when school staff are trained and when programs are aligned to the school's goals and culture (Hoelscher et al., 2004).

Similarly, research on establishing service learning programs in higher education indicates that a match between institutional mission and strategic planning, acceptance of the need for the program, and willingness to dedicate resources to support the program all facilitate program institutionalization (Bringle & Hatcher, 2000; Morton & Troppe, 1996). Finally, in his work on why educational innovations are adopted and discarded, Stevens (2004) argues that for school districts to successfully institutionalize programs, they must develop internal capacity to "monitor and maintain the program."

As programs grow to scale and are institutionalized, they present certain evaluation challenges. There is growing understanding that modern, large-scale programs intended to change the "core technology" of how students are taught (Ogawa, 2009) are incredibly complex, and the ultimate success of the program depends on the setting, players, and context (Cohen-Vogel, Tichnor-Wagner, Allen, Harrison, Kainz, Socol, & Wang, 2015, 2014; Hoenig, 2006). Research has documented features of program implementation that influence whether or not education programs "work," including leveraging the expertise of educators, providing opportunities for teachers to collaborate, engaging administrators in the implementation, and local contextual factors (Cohen-Vogel et al., 2015). An important and challenging goal of large-scale program evaluation is to determine how context matters (Honig, 2006); that is "what works, where, when, and for whom" (Means & Penuel, 2005).

In order to explore these questions of context and differential effectiveness, researchers must move beyond estimating an average treatment effect and explore heterogeneity. Often, researchers studying large-scale interventions do not have reliable measures of implementation across different sites. One approach is to estimate an overall treatment effect and then use mixed methods to understand implementation at different sites. For example, in a series of studies on a

system-wide science reform undertaken in the Los Angeles Unified School District (LAUSD), researchers explored different dimensions of the implementation of LAUSD's professional development intervention. While the central study estimated the effect of the intervention on student achievement (Borman, Gamoran, & Bowden, 2009), supplementary studies used classroom observations (Lal & Osisioma, 2009) or teacher survey data (Bruch, Grigg, & Hanselman, 2009) to document differences between treatment and comparison schools. Finally, surveys of system stakeholders (e.g., administrators, teachers) provide insight into factors that facilitate or impede the success of the intervention (Osthoff, Shewakramani, & Kelly, 2010).

The institutionalization of UA is even more complicated than many other programs. The program was started by several informal science institutions in NYC and funded by the City Council, but their target audience was teachers and students in NYC public schools. Each of the organizations had separate relationships with different sets of schools, but none had a formalized relationship with the NYCDOE. So not only did the informal organizations have to develop collaboration among themselves, they also had to develop a relationship with the DOE's central office. In previous work (Weinstein et al., 2014) we have used mixed methods to explore schools' contextual factors that influence success in UA.

In the future, as we work with the UA program to develop measures of program implementation across schools, we will explore the relationship between implementation, other contextual factors, and UA program effects. In this paper, we estimate an overall program effect and also explore heterogeneity in the impact using quantitative data. While Urban Advantage is not universally implemented in all NYC middle schools, it has operated on a large scale for more than a decade, with more than 100 schools participating in each year since 2007. Schools join and exit UA over time, providing considerable variation in the number and characteristics of

schools (and students) participating in any given year. We use this variation to explore heterogeneity in the impact of UA for different types of schools and students.

III. Data and Models

A. Data

Using data from the UA program, from the NYCDOE, and from the NYSED, we construct a longitudinal dataset of NYC middle schools and students from 2005-2014. We use program data from UA to identify participating schools in each year and to distinguish among schools based on the number of years they have participated in UA. We define UA schools as those schools that that have at least one teacher who is participating in UA in that year. Thus, as teachers move in and out of program participation, schools can be de-classified (and re-classified) as participating schools.

Student-level files from the NYCDOE student-level files provide test scores and student characteristics. Our key outcome is performance on New York State's eighth-grade Intermediate Level Science (ILS) exam, which we measure with both a z-score (standardized to have a mean of zero and a standard deviation of one) and a proficiency indicator, which takes a value of one if a student receives a three or four (out of four) on the test. We construct similar measures for performance on eighth-grade math and English language arts (ELA) exams, which we explore in robustness tests.

Student-level covariates provide demographic, program, and academic information, including race, gender, poverty (measured as eligibility for free or reduced-price lunch), limited English proficiency (LEP), eligibility for special education services (SPED), and math and English language arts (ELA) standardized test scores. We link students to their schools in every

year to determine if a student attends a UA school, and with unique student identifiers, we are able to follow students over time. Note, however, that due to data limitations we are not able to match students to their science teachers.

School Report Cards (SRCs) from the NYSED provide time-varying school characteristics, including the percent of students in the school who are black, Hispanic, Asian, poor, and limited English proficient. Additionally, SRCs include information on school size, which we measure using the natural log of total student enrollment. Our final dataset includes eighth-grade students from 2005 (the first year of UA) through 2014. In each year there are approximately 60,000-70,000 eighth-graders in 606 unique schools, for a total of 687,725 total student-year observations.

B. Main Models

We estimate a series of models to determine the effect of attending a UA school on students' performance on the eighth-grade ILS exam. Our base specification is as follows:

$$(1) \quad Science_{ijt} = \beta_0 + \beta_1 UA_{jt} + \beta_2 Zmath_{it-1} + Student_{ijt}'\beta_3 + School_{jt}'\beta_4 + \alpha_j + \gamma_t + \varepsilon_{ijt}$$

In this model, *Science* is a science outcome (either a proficiency indicator or z-score) for student *i* in school *j* in year *t*. Because proficiency is a particularly policy-relevant outcome, for most analyses we show proficiency results and include results for z-scores in the appendix.

UA is an in indicator that takes a value of one if school *j* has any teachers participating in UA in year *t*. In our preferred specification, we replace the UA indicator variable with two separate variables to distinguish between schools' first year in UA (*base year)* and all subsequent years (*post-year).* We do this because UA schools likely do not fully implement UA practices until at least their second year in the program. [2] *Post-year* is an indicator that takes a value of one

[2] Teachers in their first year of UA participate in 48 hours of professional development across three cycles throughout the year.

in all years after a school's first year in the program, regardless of whether or not it is currently participating. This is because as previously described, once teachers learn UA concepts and schools receive UA resources, the treatment cannot be fully retracted. Furthermore, exit from UA is likely endogenous; for example, struggling schools, those with high science teacher turnover, and those that have been less successful in implementing UA may be more likely to exit the program. By not "turning off" the post-year variable, we include all schools that have ever participated in UA in the treatment group.

We also control for students' lagged math z-scores. Because math and science test scores are highly correlated, these specifications will approximate value-added models, but a true value-added model is not possible, as students do not take a standardized science exam in seventh grade. Using lagged math performance, however, allows us to control for some measure of students' prior achievement.[3] *Student* is a vector of student-level characteristics, including race, gender, LEP, SPED, and poverty status. *School* is a vector of time-varying school-level characteristics, including the percent of students who qualify for free or reduced-priced lunch, the percent who are black, Hispanic, and Asian, the percent of students who are limited English proficient, and the natural log of total school enrollment.

Finally, the model includes school (α) and year effects (γ), and ε is an error term with the usual properties. Including school effects means we are comparing the performance of eighth-graders in the same school over time. That is, β_1 captures the UA program effect, comparing students in the same school before and after the school joins UA. Models using the proficiency

[3] Results, not shown, are also similar when controlling for students' lagged ELA z-scores.

indicator as the dependent variable are linear probability models, and β_1 will reflect the impact of UA on a student's likelihood of being proficient on the exam.[4]

It is important to note that we have been conservative in our definition of the UA treatment, and this leads us to potentially underestimate the impact of UA on students' science performance. Our definition of treatment is conservative in two key ways. First, as just described, we do not "turn off" the post-year variable and therefore estimate the effect of attending a school that has ever participated in UA. This is less of an issue when analyzing effects of programs that have been implemented for only a few years; as we include more years of program data, however, the likelihood that we include schools in the treatment group that have not recently participated in UA increases. This motivates our exploration of heterogeneity in the impact of UA by school characteristics, and in particular by the percent of science teachers participating in the program.

Secondly, defining the treatment at the school level will bias our estimates toward zero by including students who do not have UA teachers in the treatment group. Recall that we cannot match students to their science teachers, and therefore we estimate the impact of attending a school that has ever had any teacher participating in UA. There will be many students in the treatment group who have never had UA teachers. Furthermore, because the definition of treatment is based only on attending a UA school in eighth grade (and does not incorporate information about students' seventh-grade schools), students who attend UA schools in seventh grade but not eighth grade will be in the comparison group (non-UA schools).This issue also motivates our exploration of heterogeneity in the impact of UA, and in particular differences by student exposure to UA schools.

[4] Because we use school fixed effects, it is inappropriate to use Maximum Likelihood Estimators, such as Probit or Logit. Using either of these MLE approaches would yield inconsistent estimates (Neyman & Scott; Greene, 2004).

B. Heterogeneity of impact

We perform two key analyses to determine how the impact of UA varies by school context. First, we determine whether UA is especially effective in schools that have traditionally struggled in science. To do so, we divide schools into quartiles based on their average science z-scores in 2004, which is the year before UA was introduced in NYC.[5] We then estimate our preferred specification on these four quartiles separately. Next, we explore whether UA has a larger effect in schools with greater concentrations of science teachers participating in UA. In these schools, students have a higher likelihood of having a UA teacher, and there may be greater collaboration within the science department, more coherent curricula within and across grades, etc. To explore this in our models, we interact the UA participation variable with indicators reflecting a low concentration (0-33%), medium concentration (34-66%), high concentration (67-99%), and all teachers participating (100%). Note that schools will move in and out of the different concentration categories over time as their percent of UA teachers changes.

We also explore heterogeneity in the impact of UA by student characteristics. To determine if UA is particularly effective for certain groups of students, we estimate models on separate samples of students by race, gender, LEP, SPED, and poverty status.

C. Do UA effects grow with years of participation?

Finally, we explore how the impact of UA varies by school years of participation and by student exposure. To determine if the effect of UA grows over time within schools, we distinguish between schools' first post-year (the first year after a school's initial year), second post-year, and post-years three and beyond. Note that we do not require schools to actively participate in UA to be included in these post-year groups. For example, post-year two takes a

[5] We group schools that did not have 8th-grade test scores in 2008 into quartiles based on average science z-scores in the first year for which they are observed in our sample.

value of one if the school is in its second year *after* the base year – regardless of whether any teachers are actively participating in UA in that year.

To determine if the impact of UA varies by students' years of exposure, we estimate the following model:

$$(2)\quad Science_{ijt} = \beta_0 + \beta_1 UAorPost\_1yr_{ijt} + \beta_2 UAorPost\_2yrs_{ijt} + \beta_3 Zmath_{it-1} + Student_{ijt}'\beta_4 + School_{jt}'\beta_5 + \alpha_j + \gamma_t + \varepsilon_{ijt}$$

Here, *UAorPost_1yr* is an indicator that takes a value of one if student *i* attended a UA school (in its base year or any post-year) in either seventh grade or eighth grade; *UAorPost_2yrs* equals one if the student attended a UA school in both seventh grade and eighth grade.[6] To be clear, in this model β_1 reflects the impact of attending a UA school for only one year (either seventh or eighth grade) and β_2 reflects the impact of attending a UA school for both seventh and eighth grade, compared to students who never attend a UA school in these grades. All control variables and fixed effects are as previously defined.

While students who attend UA schools for both seventh and eighth grade may not have had UA teachers in those years, they are more likely to have had a UA teacher at some point in their middle school career than those who attend a UA school for just one year. Additionally, as previously described, some students will have attended UA schools in seventh but not eighth grade, and these students are not considered treated in the main models. Here, these students are removed from the comparison group.

IV. Results

A. Main results

[6] This group includes students who attended different UA schools in seventh and eighth grade.

Table 2 shows results for models that estimate the impact of UA on students' science proficiency (columns 1-2) and z-scores (columns 3-4). Models with the proficiency outcome are linear probability models, and coefficients reflect predicted probabilities of being proficient on the exam. In column 1, we find a small positive effect of attending a school actively participating in UA on proficiency, with students 1.5 percentage points more likely to be proficient. When distinguishing between the base year and post-years, we find a positive effect of UA in the school's first year of 2.0 percentage points, and a much larger effect of 6.0 percentage points in the post-years.

Results for z-scores also reveal a small positive relationship between UA status and student test scores, though results are less consistently significant. In column 3, we estimate a small positive coefficient on UA, but it is not statistically significant (0.010). When distinguishing between the base year and post-years in column 4, we see there is no effect of UA in the base year (0.005), but there is a small and statistically significant effect in post-year (0.028).

B. Heterogeneity in the impact of UA

The first type of heterogeneity we explore is based on schools' prior science performance. We divide schools into quartiles based on their average science z-scores in 2004 (the year before UA was introduced).[7] Quartile 1 schools have the lowest average initial science z-scores (-0.67), followed by quartile 2 schools (-0.21), quartile 3 schools (0.18), and quartile 4 schools (0.76).[8] In Table 3 we present proficiency results for our preferred specification, which

[7] For schools not in the data in 2004, we use the mean eighth-grade science z-score in the first non-UA year the school is observed in the data.

[8] In this sample, there are 115 schools in quartile 1, 145 schools in schools in quartile 2, 154 schools in quartile 3, and 139 schools in quartile 4. Figures calculated using school-level (not school-year level) data.

estimates separate effects for the base year and post-years; results for z-score outcomes are included in the appendix (Table A1).

For schools in the lowest quartile (quartile 1), we find no effect for the base year but a positive effect of UA in the post-years of 6.0 percentage points. For quartile 2 schools we estimate larger effects, with students 4.6 percentage points more likely to be proficient in the base year and 10.2 percentage points more likely to be proficient tin the post-years. In quartile 3 schools, there is no effect for the base year but a statistically significant effect in the post-years of 5.0 percentage points. Finally, for quartile 4 schools, we estimate a positive effect in the base year of 1.7 percentage points, and a larger effect of 3.5 percentage points in the post-years. Taken together, these results reveal that UA is effective at increasing proficiency in schools across the distribution of prior test scores, though effects are largest for low-performing schools (and quartile 2 schools in particular).

Next, we explore heterogeneity in the impact of UA based on the concentration of science teachers participating in UA.[9] In this model, we interact the base year and post-year indicators with indicator variables for low concentration (0-33% of teachers in UA), medium concentration (34-66%), high concentration (67-99%), and all teachers participating. Across all years in this analytic sample (2005-2014), 48.80 of schools (in either the base year or any post-year) have a low concentration of teachers participating in UA, 15.30% have a medium concentration, 13.98% have a high concentration, and 21.93% have all teachers participate.

Results for this model, shown in Table 4, reveal that effects are similar for schools in these different groups, though effects are slightly larger for schools with a high concentration of

[9] The estimated concentration of UA teachers calculated as follows: We divide the number of participating UA teachers by the estimated number of science teachers. We calculate the estimated number of science teachers by dividing total 6-8 enrollment by 135. 27 is the average middle school science class size, and we assume each science teacher has 5 classes (27*5=135).

UA teachers. Students attending a school with a low concentration of teachers in the school's base year are 1.7 percentage points more likely to be proficient on the exam, and the effect in base years is 6.0 percentage points. For schools with a medium concentration of teachers, the positive effect is 2.7 percentage points in the base year and 5.4 percentage points in the post-years. For schools with a high concentration of UA teachers, there is no effect in the base year and a larger effect in the post-years (6.8 percentage points). Finally, for schools where all teachers participate, there is no effect in the base year, though students attending in the post-years are 5.8 percentage points more likely to be proficient. Results for the z-score outcome are shown in the appendix (Table A2).

Finally, we estimate models on different student subgroups (Table 5) to determine if UA is particularly effective for certain groups of students. Compared to the overall estimated effect of attending a UA school in any post-year of 0.060 (Table 2, column 2), we estimate the largest effects for Hispanic students (0.074), male students (0.065), and non-poor students (0.062). The estimates for poor students (0.059) and female students (0.055) fall just below the overall estimate of 0.06 while point estimates are slightly smaller but still statistically significant for black (0.053), SPED (0.051), white (0.049), Asian (0.041), and LEP students (0.040). Results from subgroup models using science proficiency as the outcome are shown in the appendix (Table A3). Taken together, these results suggest that UA has a positive impact on the science outcomes of all student subgroups, though results are particularly large for Hispanic and male students.

C. Do effects vary by schools' years of participation or students' exposure?

Table 6 shows results for the model that distinguishes between schools' post-years to determine if effects grow or shrink over time. Column 1 displays results from the main

proficiency model for comparison. As shown in column 2, the impact of UA on science proficiency grows over time, as schools are in the program for more years. While attending attending a school in any post-year increases likelihood of science proficiency by 6.0 percentage points (column 1), results are somewhat smaller for schools in the first post-year (4.7 percentage points) and second post-year (5.4 percentage points). For students attending a school that is in at least its third post-year, however, the likelihood of proficiency is increased by 7.3 percentage points. Results for z-score outcomes (appendix Table A4) indicate that the impact for z-scores is similar for post-year 1, post-year 2, and post-year 3+. Taken together, these results suggest that as schools remain in UA for longer, they are better able to improve students' science proficiency, but not necessarily z-scores.

Finally, we explore the impact of repeated exposure to UA at the student level. Table 7 shows results from the model that estimates separate effects for students enrolled in UA schools for just one year (either seventh or eighth grade) and for two years (both seventh and eighth grades). Estimates reveal that proficiency results are driven by students who attend UA schools for both years. There is no impact of attending a UA school for just one year (-0.000), while students who attend a UA school for two years are 6.4 percentage points more likely to be proficient on the exam than students who never attend a UA school. Results for z-score outcomes are shown in the appendix (Table A5).

V. Summary and Conclusion

A. Summary

In this study, we use data from Urban Advantage to better understand how the impact of a program that goes from innovative to institutionalized changes over time and to address

multiple types of heterogeneity in the impact of UA. Our analysis uses ten years of data (2005-2014) to determine how middle schools' participation in UA over time affects student performance on the eighth-grade science exam. Given the size and longevity of UA in NYC, we have the unique opportunity to explore heterogeneity in the program effect by the schools' number of years participating in UA. This allows us to determine whether effects grow or shrink over time. Results from this study provide evidence about the ongoing efficacy of UA, and they more broadly provide insight into how program effects can change over time as programs become institutionalized in school districts' approach to education.

We consistently find small-to-moderate positive effects of attending a UA school on students' science performance after the school's first year of participation. Student attending UA schools in the post-years are 6.0 percentage points more likely to be proficient on the science exam and perform 0.028 SDs higher on the science exam. In our explorations of heterogeneity in the impact of UA, we find that UA is most effective in previously low-performing schools. In particular, we find large positive effects for quartile 2 schools; students who attend these schools in the post-years are 10.2 percentage points more likely to be proficient on the exam. We do not find meaningful differences in the impact of UA based on the school's concentration of science teachers participating in UA, though effects are slightly larger in schools with a high concentration of participating teachers.

In our subgroup analyses that explore heterogeneity in the UA effect based on student characteristics, we find very positive effects for all student subgroups. Effects are particularly large for Hispanic students, who are 7.4 percentage points more likely to be proficient on the exam when they attend a UA school in a post-year; results for black (5.3 percentage points), white (4.9 percentage points), and Asian students (4.1 percentage points) are also positive.

Effects are larger for male students (6.5 percentage points) than female students (5.5 percentage points), but there are not meaningful differences in the impact of UA by poverty status. Combined with the quartile results, our findings indicate that UA is having a positive impact for all the groups we analyzed, but it is especially effective for high-need schools and students. From a policy and program standpoint, this suggests UA should target the highest-need schools, potentially recruiting struggling non-UA schools directly and working to expand within active UA schools.

Results from models exploring the impact of UA by schools' years in the program indicate that effects grow over time for proficiency but not for z-scores. Finally, analysis of repeated student exposure reveals that our effects are driven by students who attend UA schools in both seventh and eighth grade.

B. Contributions and limitations

This study provides rigorous evidence about the efficacy of a large-scale science intervention, and findings are useful for policy makers and administrators of other programs interested in the development and expansion of programs over time. For example, our attention to differences in the characteristics of schools joining UA over time and differences in the impact of UA as schools remain in the program highlight important considerations for analysis of large-scale, long-term programs.

Evidence about how UA is affecting student outcomes in NYC is also useful for district-level policy makers who can help to support the program, UA program administrators who may use results to inform program development, and school principals and teachers who must determine whether or not to participate in UA. By using a rigorous empirical strategy to estimate plausibly causal program effects and to answer several nuanced questions about the impact of

UA for different groups of schools and students, we provide useful evidence about the impact of UA on eighth-grade outcomes. For example, our results suggest that UA should target the highest-need schools and students and should focus on increasing the concentration of teachers in UA schools.

Despite these meaningful contributions, this study has several limitations. First, we cannot link students to their science teachers and thus must define the UA treatment at the school level. To be clear, we define a school as participating in UA if at least one teacher in the school is participating in the program. Because the effect of UA is likely to be larger for students who both attend UA schools and have UA teachers (versus those in UA schools with non-UA teachers), defining the treatment at the school level likely leads to an under-estimate of the program effect. Thus, our estimates should be thought of as lower bounds for the true program effect.

A similar limitation is that because a school's UA status is defined based on the participation of its teachers, changes in UA status can reflect somewhat random variation in teachers' participation as opposed to strategic school-level decisions. Teacher may opt to participate or not participate in UA for a variety of reasons, including personal time constraints, other professional development opportunities, and their own perceptions of the usefulness of UA training and resources. Furthermore, as participating teachers move between schools, schools change UA status, but we cannot currently track former UA teachers as they move to new schools. This is another reason our estimates should be thought of as lower bounds.

References

Borman, G., Gamoran, A., & Bowden, J. (2009). Growing capacity or dissipation? Second-year outcomes of a school-randomized trial of the effects of professional development on student achievement in elementary science. Paper presented at the 2009 annual meeting of the American Educational Research Association, San Diego, CA.

Bringle, R. G. and Hatcher, J. A. (2000). Institutionalization of service learning in higher education. *The Journal of Higher Education*, 71(3), 273-290.

Bruch, S., Grigg, J. A., & Hanselman, P. (2009). Up to the classroom door: effects of a school-randomized trial. Paper presented at the 2009 annual meeting of the American Educational Research Association, San Diego, CA.

Cohen-Vogel, L., Tichnor-Wagner, A., Allen, D., Harrison, C., Kainz, K., Socol, A. R., and Wang, Q. (2015). Implementing educational innovations at scale: transforming researchers into continuous improvement scientists. *Educational Policy, 29*(1), 257-277.

Fisher, D. and Vinent-Lancrin, S. (2014). Measuring innovation in education, United States country note. Organisation for Economic Cooperation and Development. http://www.oecd.org/unitedstates/Measuring-Innovation-in-Education-USA.pdf

Fullan, M. and Steigelbauer, S. (1991). *The new meaning of educational change*. New York: Teachers College Press.

Greene, W. H. (2004). Fixed effects and bias due to the incidental parameters problem in the tobit model. *Econometric Reviews, 23*(2), 125-147.

Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Klein, S. P., Robyn, A., and Bugliari, D. (2003). Studying large-scale reforms of instructional practice: an example from mathematics and science. *Educational Evaluation and Policy Analysis, 25*(1), 1-29.

Hoelscher, D. M., Feldman, H. A., Johnson, C. C., Lytle, L. A., Osganian, S. K., Parcel, G. S., Kelder, S. H., Stone, E. J., and Nader, P. R. (2004). School-based health education programs can be maintained over time: results from the CATCH Institutionalization study. *Preventive Medicine,* 38, 594-606.

Honig, M. (2006). *New directions in education policy implementation: Confronting complexity*. Albany: State University of New York Press.

Klisch, Y., Miller, L. M., Wang, S., & Epstein, J. (2012). The impact of a science education game on students' learning and perception of inhalants as body pollutants. *Journal of Science Education and Technology, 21*(2), 295-303.

Lal, S. R. & Osisioma, I. (2009). The classroom perspective: exploring the complexities and challenges of implementing a fourth-grade science immersion curriculum. Paper presented at the 2009 annual meeting of the American Educational Research Association, San Diego, CA.

Means, B., & Penuel, W. R. (2005). Research to support scaling up technology based innovations. In C. Dede, J. Honan, & L. Peters (Eds.), *Scaling up success: lessons from technology-based educational improvement* (pp. 176-197). NewYork, NY: Jossey-Bass.

Morton, K. and Troppe, M. (1996). From the margin to the mainstream: Campus Compact's project on integrating service with academic study. *Journal of Business Ethics,* 15, 21-32.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1-32.

Ogawa,R. T. (2009). Improvement or reinvention: Two policy approaches to school reform. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 534-549). New York, NY: Routledge.

Osthoff, E. J., Shewakramani, V., & Kelly, K. (2010). System-side reform in science: the impact of district and school context. SCER Working Paper No. 2010-4.

Steckler, A. and Goodman, R. M. (1989). How to institutionalize health promotion programs. *American Journal of Health Promotion,* 3(4), 34-43.

Stevens, R. J. (2004). Why do educational innovations come and go? What do we know? What can we do? *Teaching and Teacher Education*, 20, 389-396.

U.S. Department of Education (2015). Update on the Investing in Innovation (i3) Fund. Last modified 01/12/2015.http://www2.ed.gov/programs/innovation/index.html?utm_source=rssutm_medium=rrssutm_campaign=the-u-s-department-of-education-announced-the-start-of-the-134-million-2014-investing-in-innovation-i3-grant-competition#update

Weinstein, M.,Whitesell, E. R., Leardo, M., Grajo, G., and Saldivia, S. (2014). Successful schools: how school-level factors influence success with Urban Advantage. IESP Working Paper No. 01-14.

Weinstein, M., Whitesell, Emilyn R., & Schwartz, A. E. (2014). Museums, zoos, and gardens: How formal-informal partnerships can impact urban students' performance in science. *Evaluation Review,* 38(6), 514-545.

Figures and Tables

Figure 1: Urban Advantage program logic model

Activities	Outcomes	Goals

• Professional development for teachers, school administrators, and parent coordinators
• Students completing long-term science investigations (exit projects)
• Access to and resources provided by science-rich cultural institutions for students and teachers
• Leadership Institutes for school-based science leadership teams and lead science teachers
• Outreach to families by science-rich cultural institutions
	• **Student Outcomes**
	  • Improved quality of long-term student science investigations (exit projects)
	  • Increased proficiency on New York State Intermediate Level Science assessment
	  • Increased enrollment in STEM high schools
	  • Greater success on high school Regents science exams
	• **Teacher Outcomes**
	  • Greater implementation fidelity of inquiry-based instructional practices
	  • Ongoing use of formative assessments to inform progress in students' science learning
		• Improve students' middle school science achievement in order to increase participation and success in high school science courses that lead to greater college readiness
		• Increase participation of high-need students in inquiry-based science learning experiences that incorporate rigorous and relevant project-based contextual learning opportunities
		• Improve teacher practice through the use of inquiry-based instructional strategies and performance-based formative assessments
		• Inform new models of STEM-focused middle school designs

Figure 2: Number and percent of UA schools by year, 2005-2014

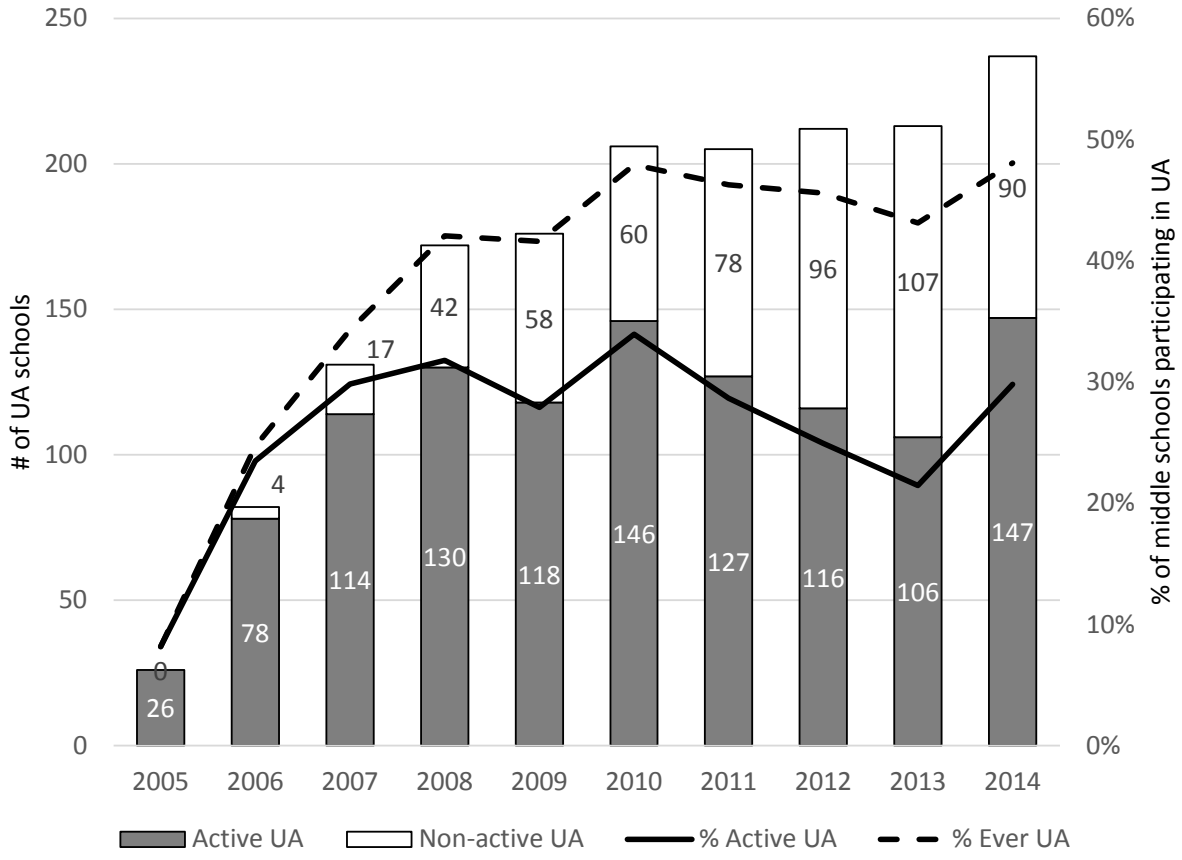
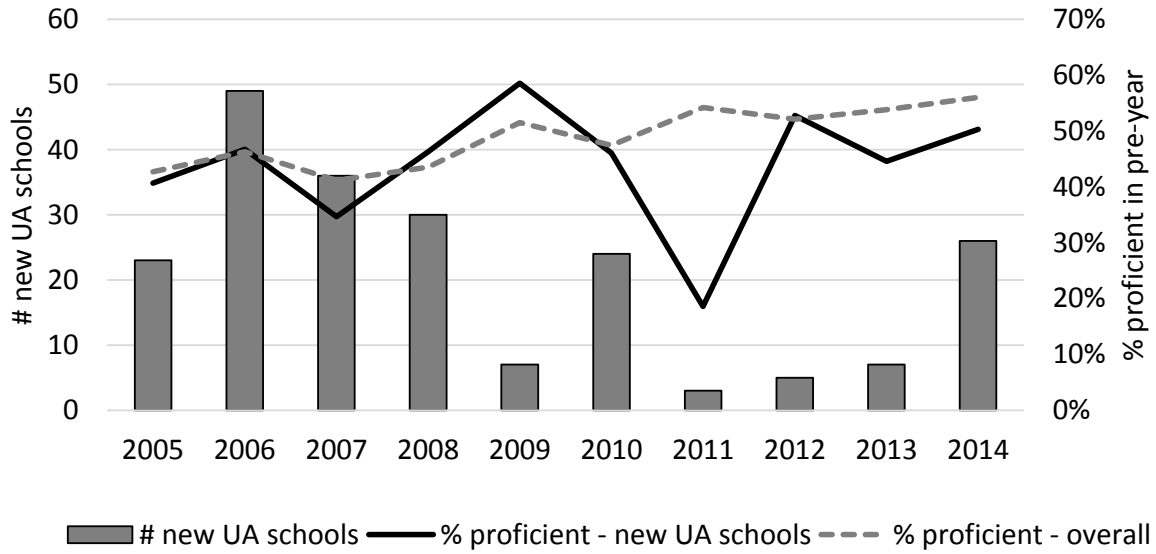


Figure 3: Prior performance of new UA schools, by year; 2005-2013

A. Percent of eighth-graders who are proficient in science at new UA schools in the year before joining UA



B. Average science z-score of eighth-graders of new UA schools in the pre-year

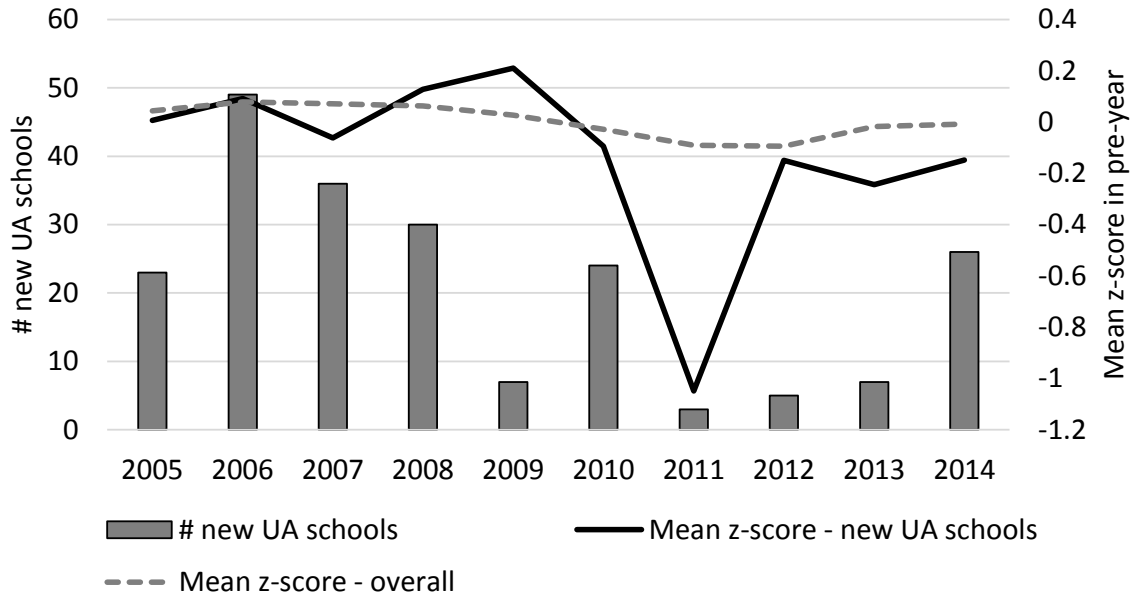


Table 1: 8th-grade student characteristics and school characteristics for UA and non-UA schools, 2005 and 2014

	2005			2014		
	UA	Not UA	p-value of difference	UA	Not UA	p-value of difference
Total enrollment	**1178**	**828**	**0.00**	**730**	**591**	**0.00**
Student demographics (%)						
Black	38.5%	35.6%	0.77	32.7%	40.5%	0.10
Hispanic	50.0%	41.8%	0.42	42.2%	39.3%	0.55
Asian	7.7%	7.9%	0.97	**17.0%**	**9.0%**	**0.01**
White	3.9%	14.7%	0.12	8.2%	11.3%	0.30
Poor	80.8%	84.9%	0.57	80.3%	78.0%	0.58
LEP	15.4%	7.2%	0.14	11.6%	8.4%	0.27
SPED	11.5%	7.2%	0.42	16.3%	16.5%	0.42
% Proficient						
Math	41.7%	42.0%	0.98	23.0%	21.9%	0.80
Reading	34.8%	35.1%	0.98	26.8%	27.5%	0.87
Science	45.5%	43.0%	0.83	51.0%	50.4%	0.91
School location (%)						
Manhattan	23.1%	20.5%	0.76	20.4%	22.0%	0.70
Bronx	15.4%	25.0%	0.27	23.8%	26.0%	0.61
Brooklyn	34.6%	34.9%	0.97	31.3%	31.2%	0.99
Queens	23.1%	16.4%	0.39	21.1%	17.9%	0.41
Staten Island	3.8%	3.1%	0.83	3.4%	2.9%	0.76
N	26	292		147	346	

Note: Bold indicates differences are statistically significant at the 0.05 level or less. The percent proficient for math, reading, and science exams is the percent of students scoring in levels 3 or 4 (out of 4) on these exams. Characteristics and proficiency rates are for eighth-grade students only. Total enrollment includes *all students* in the schools.

Table 2: Impact of UA on 8th-grade students' science outcomes, 2005-2014

	Proficiency		Z-score	
	(1)	(2)	(3)	(4)
UA	0.015***		0.010	
	(0.005)		(0.009)	
Base year		0.020***		0.005
		(0.006)		(0.013)
Any post-year		0.060***		0.028**
		(0.006)		(0.012)
Lagged math z-score	0.225***	0.225***	0.541***	0.541***
	(0.002)	(0.002)	(0.004)	(0.004)
Black	-0.093***	-0.093***	-0.201***	-0.201***
	(0.003)	(0.003)	(0.005)	(0.005)
Hispanic	-0.052***	-0.052***	-0.112***	-0.112***
	(0.003)	(0.003)	(0.005)	(0.005)
Asian	0.003	0.003	0.026***	0.026***
	(0.003)	(0.003)	(0.005)	(0.005)
Female	-0.033***	-0.033***	-0.050***	-0.050***
	(0.001)	(0.001)	(0.002)	(0.002)
LEP	-0.199***	-0.199***	-0.498***	-0.498***
	(0.004)	(0.004)	(0.009)	(0.009)
SPED	-0.160***	-0.160***	-0.337***	-0.337***
	(0.003)	(0.003)	(0.005)	(0.005)
Poor	-0.018***	-0.018***	-0.046***	-0.046***
	(0.002)	(0.002)	(0.004)	(0.004)
Schl char	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
School FE	Y	Y	Y	Y
Observations	626,920	626,920	626,920	626,920
R-squared	0.370	0.371	0.532	0.532

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Table 3: Impact of UA on 8th-grade students' science proficiency, by initial quartile of mean science z-score, 2005-2014

	Q1	Q2	Q3	Q4
Base year	0.021	0.046***	0.006	0.017*
	(0.016)	(0.014)	(0.011)	(0.010)
Any post-year	0.060***	0.102***	0.050***	0.035***
	(0.019)	(0.014)	(0.009)	(0.009)
Student characteristics	Y	Y	Y	Y
School characteristics	Y	Y	Y	Y
Year fixed effects	Y	Y	Y	Y
School fixed effects	Y	Y	Y	Y
Observations	80,031	146,878	209,672	166,679
R-squared	0.221	0.271	0.327	0.299

Robust standard errors in parentheses*** p<0.01, ** p<0.05, * p<0.1

Note: Initial quartiles are based on average eighth-grade science test scores in 2004 (the year before UA was started) or in the first year the school is observed in the data. UA schools that are not observed before joining UA are not included. Sample excludes students who are in UA for only one year (either 7th or 8th grade). Sample excludes students who are in UA for only one year (either 7th or 8th grade). Quartile 1 includes 169 schools, with initial mean z-scores ranging from -2.53 to -0.41 (mean -0.82). Quartile 2 includes 168 schools with initial mean z-scores ranging from -0.40 to -0.03 (mean -0.22). Quartile 3 includes 174 schools with initial mean z-scores ranging from -0.025 to 0.41 (mean 0.19). Finally, Quartile 4 includes 173 schools with initial mean z-scores ranging from 0.41to 1.65 (mean 0.78). Time-varying student characteristics include lagged math z-score , black, Hispanic, Asian, female, LEP, SPED, and poor. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Table 4: Impact of UA on students' science proficiency by concentration of teachers in UA, 2005-2014

	(1)
Low concentration (0-33%)	
Base year	0.017**
	(0.008)
Any post-year	0.060***
	(0.007)
Medium concentration (34-66%)	
Base year	0.027**
	(0.011)
Any post-year	0.054***
	(0.007)
High concentration (67-99%)	
Base year	0.016
	(0.015)
Any post-year	0.068***
	(0.009)
All teachers (100%)	
Base year	0.026
	(0.017)
Any post-year	0.058***
	(0.009)
Student characteristics	Y
School characteristics	Y
Year fixed effects	Y
School fixed effects	Y
Observations	626,920
R-squared	0.371

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Teacher concentration is based on the estimated percent of science teachers in the school who participate in UA and ranges from 0-100. Time-varying student characteristics include lagged math z-score, black, Hispanic, Asian, female, LEP, SPED, and poor. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Table 5: Impact of UA on 8th-grade students' science proficiency, by subgroup, 2005-2014

	(1) Black	(2) Hispanic	(3) Asian	(4) White	(5) Male	(6) Female	(7) LEP	(8) SPED	(9) Poor	(10) Not poor
Base year	0.018**	0.027***	-0.006	0.035***	0.021***	0.020***	0.006	0.021**	0.018***	0.033***
	(0.009)	(0.008)	(0.010)	(0.011)	(0.007)	(0.007)	(0.012)	(0.009)	(0.007)	(0.010)
Any post-year	0.053***	0.074***	0.041***	0.049***	0.065***	0.055***	0.040***	0.051***	0.059***	0.062***
	(0.009)	(0.007)	(0.008)	(0.010)	(0.006)	(0.007)	(0.010)	(0.009)	(0.006)	(0.010)
Student characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
School characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
School fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	191,572	250,703	96,703	87,650	313,486	313,434	66,751	67,116	512,803	114,117
R-squared	0.309	0.327	0.354	0.336	0.369	0.377	0.210	0.247	0.357	0.394

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Time-varying student characteristics include lagged math z-score, black, Hispanic, Asian, female, LEP, SPED, and poor. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Table 6: Impact of UA on 8th-grade students' science proficiency, by school years of participation, 2005-2014

	(1)	(2)
Base year	0.020***	0.022***
	(0.006)	(0.006)
Any post-year	0.060***	
	(0.006)	
Post-year 1		0.047***
		(0.007)
Post-year 2		0.054***
		(0.007)
Post-year 3+		0.073***
		(0.006)
Student characteristics	Y	Y
School characteristics	Y	Y
Year fixed effects	Y	Y
School fixed effects	Y	Y
Observations	626,920	626,920
R-squared	0.371	0.371

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Base year takes a value of one in the school's first year participating in UA. Post-year is an indicator taking a value of one in all subsequent years (does not "turn off"). Post-years 1, 2, and 3+ are mutually exclusive. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Table 7: Impact of repeated exposure to UA on 8th-grade students' science proficiency, 2005-2014

	(1)
UA or post for 1 year	-0.000
	(0.005)
UA or post for 2 years	0.064***
	(0.005)
Student characteristics	Y
School characteristics	Y
Year fixed effects	Y
School fixed effects	Y
Observations	626,920
R-squared	0.372

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Base year takes a value of one in the school's first year participating in UA. Post-year is an indicator taking a value of one in all subsequent years (does not "turn off"). Post-years 1, 2, and 3+ are mutually exclusive. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Appendix

Table A1: Impact of UA on 8th-grade students' science z-scores, by initial quartile of mean science z-score, 2005-2014

	Q1	Q2	Q3	Q4
Base year	0.036	0.063**	-0.025	-0.011
	(0.037)	(0.029)	(0.022)	(0.022)
Any post-year	0.071	0.118***	0.004	-0.043**
	(0.044)	(0.029)	(0.018)	(0.018)
Student characteristics	Y	Y	Y	Y
School characteristics	Y	Y	Y	Y
Year fixed effects	Y	Y	Y	Y
School fixed effects	Y	Y	Y	Y
Observations	80,031	146,878	209,672	166,679
R-squared	0.338	0.396	0.480	0.511

Robust standard errors in parentheses*** p<0.01, ** p<0.05, * p<0.1

Note: Initial quartiles are based on average eighth-grade science test scores in 2004 (the year before UA was started) or in the first year the school is observed in the data. UA schools that are not observed before joining UA are not included. Sample excludes students who are in UA for only one year (either 7th or 8th grade). Quartile 1 includes 169 schools, with initial mean z-scores ranging from -2.53 to -0.41 (mean -0.82). Quartile 2 includes 168 schools with initial mean z-scores ranging from -0.40 to -0.03 (mean -0.22). Quartile 3 includes 174 schools with initial mean z-scores ranging from -0.025 to 0.41 (mean 0.19). Finally, Quartile 4 includes 173 schools with initial mean z-scores ranging from 0.41to 1.65 (mean 0.78). Time-varying student characteristics include lagged math z-score , black, Hispanic, Asian, female, LEP, SPED, and poor. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Table A2: Impact of UA on students' science z-scores by concentration of teachers in UA, 2005-2014

	(1)
Low concentration (0-33%)	
Base year	-0.001
	(0.016)
Any post-year	0.027**
	(0.013)
Medium concentration (34-66%)	
Base year	0.020
	(0.021)
Any post-year	0.019
	(0.016)
High concentration (67-99%)	
Base year	-0.008
	(0.030)
Any post-year	0.042**
	(0.018)
All teachers (100%)	
Base year	0.017
	(0.036)
Any post-year	0.032
	(0.020)
Student characteristics	Y
School characteristics	Y
Year fixed effects	Y
School fixed effects	Y
Observations	626,920
R-squared	0.532

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Teacher concentration is based on the estimated percent of science teachers in the school who participate in UA and ranges from 0-100. Time-varying student characteristics include lagged math z-score, black, Hispanic, Asian, female, LEP, SPED, and poor. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Table A3: Impact of UA on 8th-grade students' science z-scores, by subgroup, 2005-2014

	(1) Black	(2) Hispanic	(3) Asian	(4) White	(5) Male	(6) Female	(7) LEP	(8) SPED	(9) Poor	(10) Not poor
Base year	0.017	0.013	-0.048**	0.027	0.006	0.005	-0.030	0.022	0.002	0.015
	(0.020)	(0.016)	(0.020)	(0.021)	(0.014)	(0.013)	(0.030)	(0.020)	(0.014)	(0.020)
Any post-year	0.040**	0.048***	-0.021	-0.002	0.040***	0.015	-0.009	0.050***	0.027**	0.023
	(0.020)	(0.014)	(0.018)	(0.019)	(0.013)	(0.012)	(0.026)	(0.019)	(0.012)	(0.019)
Student characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
School characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
School fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	191,572	250,703	96,703	87,650	313,486	313,434	66,751	67,116	512,803	114,117
R-squared	0.443	0.469	0.543	0.533	0.525	0.545	0.302	0.379	0.512	0.573

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Time-varying student characteristics include lagged math z-score, black, Hispanic, Asian, female, LEP, SPED, and poor. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Table A4: Impact of UA on 8th-grade students' science z-scores, by school years of participation, 2005-2014

	(1)	(2)
Base year	0.005	0.005
	(0.013)	(0.013)
Any post-year	0.028**	
	(0.012)	
Post-year 1		0.029*
		(0.015)
Post-year 2		0.027*
		(0.015)
Post-year 3+		0.027**
		(0.013)
Student characteristics	Y	Y
School characteristics	Y	Y
Year fixed effects	Y	Y
School fixed effects	Y	Y
Observations	626,920	626,920
R-squared	0.532	0.532

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Base year takes a value of one in the school's first year participating in UA. Post-year is an indicator taking a value of one in all subsequent years (does not "turn off"). Post-years 1, 2, and 3+ are mutually exclusive. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors are clustered at the school-year level.

Table A5: Impact of repeated exposure to UA on 8th-grade students' science proficiency, 2005-2014

	(1)
UA or post for 1 year	-0.061***
	(0.011)
UA or post for 2 years	0.046***
	(0.011)
Student characteristics	Y
School characteristics	Y
Year fixed effects	Y
School fixed effects	Y
Observations	626,920
R-squared	0.533

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Time-varying student characteristics include lagged math z-score (math results) or lagged ELA z-scores (ELA results), black, Hispanic, Asian, female, LEP, SPED, and poor. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors clustered at the school-year level.

Table A6: Impact of UA on 8th-grade students' math and ELA z-scores, 2005-2014

	Proficiency		Z-score	
	Math	ELA	Math	ELA
Base year	-0.016**	-0.011**	0.024*	0.027**
	(0.006)	(0.005)	(0.013)	(0.012)
Any post-year	-0.046***	-0.027***	0.046***	0.063***
	(0.006)	(0.004)	(0.012)	(0.011)
Student characteristics	Y	Y	Y	Y
School characteristics	Y	Y	Y	Y
Year fixed effects	Y	Y	Y	Y
School fixed effects	Y	Y	Y	Y
Observations	649,793	641,524	649,793	641,524
R-squared	0.428	0.390	0.572	0.526

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Time-varying student characteristics include lagged math z-score (math results) or lagged ELA z-scores (ELA results), black, Hispanic, Asian, female, LEP, SPED, and poor. Time-varying school characteristics include % poor % black, % Hispanic, % Asian, % LEP, and log of total school enrollment. Constant not shown. Robust standard errors clustered at the school-year level.