The Urban Advantage: Comprehensive Science Professional Development and Student Achievement

Meryle Weinstein[1], Menbere Shiferaw[2], Kaitlyn G. O'Hagan[3]

[1] Steinhardt School of Culture, Education, and Human Development, New York University

[2] Mathematica

[3] Wagner School of Public Service, New York University

Author Note

Meryle Weinstein https://orcid.org/0000-0001-6564-7196

Menbere Shiferaw https://orcid.org/0000-0001-5401-6786

Kaitlyn G. O'Hagan https://orcid.org/0000-0002-7292-7361

We thank the Urban Advantage staff at the Gottesman Center for Science Teaching and Learning at the American Museum of Natural History and the NYC Department of Education Research and Policy Support Group for their help with this study. We also thank participants at the NYU education summer seminar series for valuable feedback. Major public support for Urban Advantage is provided by the Speaker and the City Council of New York and the New York City Department of Education. This research is supported by the Institute of Education Sciences (IES), U.S. Department of Education, through Grant R305B080019 to New York University. Kaitlyn O'Hagan was also supported by IES Grant #R305B200010. In addition, funds for this study, for all authors, were received from the American Museum of Natural History.

Correspondence concerning this article should be addressed to Meryle Weinstein. Email: meryle.weinstein@nyu.edu.

Abstract

This study evaluates the impact of the Urban Advantage (UA) program on eighth grade science test scores. UA is a collaboration between the New York City (NYC) Department of Education and eight NYC informal science education institutions that began in 2004 and currently serves two-thirds of all NYC middle schools. The UA program provides high-intensity teacher professional development and additional support services (e.g. field trips, materials, principal engagement). We contribute to the literature on science professional development interventions to improve student outcomes, using a standardized assessment to assess impact and seven years of student-, school- and teacher-level data. We capitalize on the availability of unique student-teacher linkage and course data that allows us to identify students who have a UA teacher for science, unlike an earlier study, which defined the treatment at the school level. Our empirical strategy relies on matching to create a treatment and comparison group with similar observed characteristics. Results suggest, across *all* schools, performance in eighth grade science is not higher for students taught by a UA teacher compared to those without a UA teacher. However, comparing students within the same school, students with a UA teacher perform 0.02 standard deviations higher than students without a UA teacher. The magnitude of effects differs across subgroups; for instance, we find students with disabilities with a UA teacher are 1.5 pp more likely to meet eighth grade science proficiency standards compared to similar students in the same school. The analyses provide evidence that UA continues to be a successful intervention—though impacts may be smaller than previously estimated. In context with prior research on professional learning in science, the positive findings suggest districts with comparable access to informal science education institutions may want to implement similar programming.

Keywords: professional development, science achievement, middle school

The Urban Advantage: Comprehensive Science Professional Development and Student Achievement

Can collaborations between public schools and informal science education institutions (ISEIs) improve science education? Educators and researchers have documented the importance of informal learning spaces for both students and teachers to enhance students' interest in, engagement with, and understanding of science (Allen & Crowley, 2014; Ash & Lombana, 2012). While over 70% of ISEIs in the U.S. have programs specifically designed for schools, few have been institutionalized within a school system (Philips *et al.*, 2007; Bevan *et al.*, 2010). Science, technology, engineering, and math (STEM) play a critical role in the nation's economy and individuals' career opportunities, and these collaborations may help take steps towards improving STEM skills for U.S. students. However, there remains a lack of strong, empirical evidence of the benefits of IESIs for improving student achievement.

This study presents results on the impact of the Urban Advantage (UA) program on eighth grade science test scores. UA is a large-scale collaboration between the New York City (NYC) Department of Education and eight NYC IESIs that began in 2004 with 31 middle schools and over the past 19 years has expanded to reach two-thirds of all NYC middle schools (and some elementary schools). In this analysis, we capitalize on the availability of unique student-teacher linkage and course data that allows us to identify students who have a UA teacher for science, that is, those students actually receiving the treatment, unlike an earlier study (Weinstein *et al.*, 2014), which defined the treatment at the school level and likely conflated both students who were and were not actually taught by a UA teacher. The prior study found that participation in UA led to modest improvements in students' performance on New York State's (NYS) eighth grade science exam; students at UA schools outperformed students at non-UA

schools by approximately 0.05 standard deviations, with larger effects for Black students, male students, and students in special education. However, because treatment included students who were not actually taught by a UA teacher, these estimates are likely biased.

With the new student-teacher linkage data, we are able to explore if there are differences in students' performance in eighth grade science for students who were and were not taught by a UA teacher. With these data, we are also able to account for important teacher-level confounders that can influence student performance and program participation, such as years of teaching experience, a crucial predictor of student performance (Ladd & Sorensen, 2017; Harris & Sass, 2011). Our empirical strategy relies on matching to create a treatment and comparison group with similar student, teacher, and school characteristics.

The results show that overall, performance in eighth grade science is not higher for students taught by a UA teacher compared to those without a UA teacher across all schools. However, in models comparing students within the same school, students with a UA teacher performed roughly 0.02 standard deviations higher than students without a UA teacher. Though these differences in test scores generally do not translate into differences in meeting proficiency standards, there are different patterns in the magnitude of effects for subgroups of students. For instance, we find students with disabilities (SWD) with a UA teacher are 1.5 pp more likely to meet eighth grade science proficiency standards, and English language learners (ELL) with a UA teacher are 1.9 pp more likely to meet eighth grade science proficiency standards, compared to similar students in the same school. The analyses presented in this paper provide evidence that UA continues to be successful as a school-level intervention—though impacts may be smaller than previously estimated. Regardless, in context with prior research, these positive findings suggest other districts with comparable access to informal science education institutions may

want to implement similar programming.

The Urban Advantage (UA) Program

UA is designed to provide teachers and students in NYC middle schools (schools serving grades 6-8) the opportunity to engage in authentic science practice through professional development for teachers, classroom materials, administrative support, outreach to families, and access to cultural institutions. The professional development (PD) model provides intense, ongoing, and authentic hands-on learning experiences for teachers. The PD takes place at participating ISEIs and is conducted by UA program staff, UA lead teachers (experienced UA teachers who support other UA teachers in their school), and informal science educators from the host institutions. The PD model as designed also brings together the broader scientific community and includes geologists, astronomers, and biologists. As part of their training, teachers conduct their own scientific investigations, and experience firsthand what it means to "do science," consistent with the teacher-as-learner model of PD, which has proven to be effective for teachers in STEM education (Loucks-Horsley & Matsumoto, 1999; Loucks-Horsley *et al.*, 2010; National Research Council, 2009; Thompson & Zeuli, 1999). PD sessions are available at all the participating institutions and held on weekends, district-wide PD days, or during the week.

The program is designed to meet the needs of both novice and experienced teachers. UA has created levels of professional learning, such that teachers have access to offerings that match their level of experience. During their first year, teachers attend 40 hours of professional learning focused on scientific investigations and the science-rich cultural institutions themselves. Teachers in years two and three complete 22.5 hours of PD which target teachers' abilities to *support students* in conducting long-term scientific investigations. Teachers are also trained to

use a variety of classroom tools developed by UA staff intended to support them as they apply UA principles in the classroom. Teachers who have been in the program for more than three years complete 12.5 hours of continuing professional learning each year. The highest levels of professional learning culminate in offerings around 'reflective practice,' in which teachers bring student work or videos of their own teaching to analyze. In addition, participating teachers receive funds to purchase materials to use in their classrooms. UA teachers, administrators, students, and families also receive vouchers for free admission to any of the ISEIs and schools receive transportation funds to facilitate these trips.

Literature Review: Effect of Programs Similar to UA on Student Outcomes

Numerous science professional development interventions have included inquiry-based science, incorporated working scientists to lead the professional development sessions, and/or incorporated field trips to ISEIs for students and teachers. However, few, if any have combined all of these practices into one program or brought together all of the resources available in one city with a large public school system, as the Urban Advantage program does (Loucks-Horsley *et al.*, 2010).

Teaching and learning at ISEIs differs from school-based learning in a number of ways. It takes place at a variety of venues, such as museums, parks, zoos, and gardens; it focuses on learners' choices; and includes structures that incorporate the learners' motivations, culture, and ability (National Research Council, 2009). Additionally, unlike schools, ISEIs do not conduct formal, high-stake assessments. While most of these institutions have provided learning opportunities to children, youth, and adults, few have been working alongside public schools to improve science achievement among students (National Research Council, 2009).

There are many programs that have been implemented to improve science teaching and

learning, yet only a few studies exist that provide evidence of their effectiveness in improving student achievement in science. Additionally, most of these programs have been adopted by a few schools in one district or across districts, and almost none of these programs have been implemented at-scale across a school district, as the UA program has been. Furthermore, few use rigorous study designs or high-stakes standardized tests to measure impact, instead relying on assessments created by the program developers to show differences in achievement among those in the program and those who did not participate.

In a 2007 review of the evidence on teacher professional development and student achievement, Yoon and colleagues found that only nine of 1,300 studies identified met the What Works Clearinghouse evidence standards (Yoon *et al.*, 2007). All of these focused on elementary school age students and teachers and of those, only two focused on science. Even ten years later, there were only a handful of studies that focused on professional development for secondary school science teachers and used rigorous methods to examine impact: A 2017 best-evidence synthesis (Cheung *et al.*, 2017) focused exclusively on programs in grades 6-12. They found 21 studies that used either quasi-experimental methods or random assignment that qualified for their analysis, and six studies of instructional process programs, which "provided substantial professional development and coaching to teachers in specific approaches to inquiry-oriented science teaching." The average effect size of these programs was 0.17 sd, although across programs it ranged from 0.07 to 0.46 sd, and only three of the programs used state standardized tests to evaluate impacts (the other three used program or curriculum-specific assessments). While these "instructional process programs", of those reviewed, are the most similar to the UA program, none of these programs linked PD and science-rich cultural institutions. Indeed, there are studies of programs similar to UA, in terms of offering *science* professional development to

teachers (including those reviewed by Cheung *et al.*), partnerships between schools and outside educational institutions, and/or field trips/site visits for students—but not all of these components. However, we review some of this evidence to benchmark potential effect sizes of a program such as UA.

Four recent studies (not included in the reviews discussed above, given they were published more recently) used an experimental design (random treatment assignment with a control group) to examine the impact of science professional development on student achievement (Harris *et al.*, 2022; Krajcik *et al.*, 2022; Schneider *et al.*, 2022; Zoblotsky *et al.*, 2017).

Harris *et al.* (2022) evaluated the impact of implementing the Amplify Science Middle School curriculum on 1,780 seventh grade students in 15 schools across three districts. The curriculum came with 24 hours of professional learning for teachers provided by Berkeley's Lawrence Hall of Science. They found that students in the treatment group scored 7.3% higher than the comparison group on a physical science assessment developed by the research team, and impacts were similar across gender and racial and ethnic groups and for students with different prior math and literacy achievement.

Krajcik *et al.* (2022) examined the impact of a science intervention on 2,371 third grade students across 46 schools in Michigan. The professional learning for teachers focused on implementing project-based learning. Students in the treatment group performed 0.28 standard deviations (sd) higher than those in the control group on the science component of Michigan's state standardized test.

Schneider *et al.* (2022) examined the impact of science PD on 4,237 high school students studying physics and chemistry across 61 schools in California and Michigan. The professional

learning for teachers, which focused on a project-based learning in physical science, was developed by personnel at the Lawrence Hall of Science and emphasized teachers' active participation in learning, connections to classroom contexts, collaboration, and reflection. Students in the treatment group performed 0.20 standard deviations (sd) higher than those in the control group on an independently developed assessment; similar effects were found across race/ethnicity and gender subgroups.

Zoblotsky *et al.* (2017) evaluated the impact of *Leadership and Assistance for Science Education Reform* (LASER), a program developed by the Smithsonian Science Education Center that provides professional development and science kits to participating teachers. The evaluation considered the impact of LASER on science achievement of elementary and middle school students schools in three states. Students were assessed using the WestEd-developed Partnership for Standards-based Science Assessment (PASS). The middle school study, which considered approximately 2,200 students across 11 schools, found no statistically significant impacts on science achievement.

On additional study, Seraphin *et al.* (2017), examined the impact of a professional development program for aquatic science, but used a pre/post single group design. The program provided both in-person and online professional development to 27 teachers from all grade levels in Hawaii. Using a test designed for the study, rather than a standardized assessment, the authors found a change in student knowledge ranging from 0.1 sd to 0.3 sd.

Considering these studies and studies covered in prior reviews, the preponderance of evidence suggests science programs that include a significant professional development program have positive effects—while some studies found small or null impacts, others found impacts of 0.2-0.4 sd. It is unclear if the differences in impacts are due to actual difference in program

impacts (and therefore, effectiveness or underlying program components), or the method of assessment. Using state standardized tests to evaluate impact allows for a broader comparison group (since most students take these exams, regardless of whether they receive the program of interest), and may better reflect the ability of a program to impact outcomes at scale, and on science learning more broadly.

Looking beyond studies of PD-intensive science programs, a few studies have focused on whether programs in partnership with cultural institutions though field trips and other types of experiences improve student learning outcomes. For example, Lacoe *et al.* (2020) examined the impact of a long-standing museum-based educational program for low-income elementary school students in San Diego. Using a difference-in-difference design and standardized tests, they found 0.01 sd increases in math and 0.07 sd ELA in the year of participation. Greene *et al.* (2014) studied the impact of field trips to the Crystal Bridges Museum of American Art using a randomized control design. The results show that students in the treatment group had a stronger ability to think critically, had a higher level of historical empathy, and increased tolerance compared to the control group, but they did not look at impacts on standardized tests. Whitesell (2016) also studied the impact of field trips and found small positive effects of exposure to field trips on students' science test scores. Using data from the Urban Advantage program from 2007-2012, she estimated positive impacts of exposure to field trips on student's scores on state standardized science exams: for each additional class visit, student scores increased by 0.018 sd. While this average result was not statistically significant, among students with high exposure to field trips, the impact was 0.26 sd and statistically significant.

None of the programs evaluated (except for Whitesell, 2016, which studied the field trip component of the UA program) incorporates all of the best practices implemented by UA: high-

intensity ongoing teacher professional development *and* partnership with outside science education institutions (including both field trips for students and on-site PD with scientists for the teachers). In addition, few of the evaluations had a sample that was large enough to conduct subgroup analyses, or used standardized assessments to assess impact. Finally, most previous evaluations have used only one or a few years of data, while we are able to draw on seven years of data to evaluate impact. Thus, this study of UA adds to the literature on the impact of programs to improve student outcomes through professional learning, with a particular focus on middle school science.

Methods

Data and Sample

This analysis uses detailed student-, teacher-, and course-level data provided by the NYC Department of Education (NYC DOE), from school years 2013-2019, in addition to school-level data from the New York State *School Report Cards*. The NYC DOE data have unique person and school identifiers that allow us to track both teachers and students across schools and over time.[1] These data include teacher-student linkage files, student demographic, educational and test score files, teacher personnel files, and a UA program file that is a teacher-level dataset that identifies teachers participating in UA and their school in each year. The teacher-student linkage files identify students, their teachers, and the corresponding course in which the teacher has the student. The student-level files include sociodemographic characteristics (gender, race/ethnicity, eligibility for free/reduced price lunch), educational needs (student with disability, English language learner), and standardized test scores in English Language Arts (ELA) and mathematics

[1] UA staff provided the NYCDOE with a list of participating teachers by year who then matched these teachers to a scrambled teacher identification number, which was then provided to the researchers and allows us to track teachers over time. NYCDOE similarly provides scrambled student identification numbers.

in grades 3 to 8, and science in grades 4 and 8. The teacher personnel data contain teaching assignment, licensing, and the number of years teaching at the school and with the NYCDOE. Taken together, these data allow us to identify which students are taught by UA teachers and control for student-, teacher-, and school-level characteristics that may be associated with both selection into UA and science achievement, reducing bias in our effect estimates. We limit our sample to the years 2013-2019 because no teacher-student match data is available prior to 2013 and achievement data is only available through 2019 because of the COVID-19 pandemic.

Our outcome, science achievement, is measured using the eighth grade Intermediate Level Science (ILS) exam in two ways. Principally, we measure performance with a standardized score (z-score), which is a measure of relative performance standardized across students within a grade and year to have a mean of 0 and standard deviation of 1. Students performing above (below) average relative to other students in their grade, in that year, have positive (negative) z-scores. As a second measure, we evaluate the probability that students meet the ILS exam proficiency benchmarks to gauge whether exam score improvements translate into meeting the standards. New York State divides the scale scores into performance levels 1 through 4, and students who score in levels 3 or 4 are designated by the state as meeting the standards.

The sample contains eighth grade students who took the ILS exam and who could be matched to their science teacher. The student-teacher linkage data allow us to match students with their science teacher in each academic year. Thus, we are able to identify students that have a UA science teacher and those who do not. We use three exclusion criteria to arrive at our analysis sample. We exclude charter schools because they lack student-linkage data. We exclude special education only schools (District 75 in NYC) that educate students with severe disabilities

since they have few tested students or UA teachers. Lastly, we exclude teachers and schools with fewer than 10 students who took the eighth grade ILS exam (the outcome).

Our analysis focuses on students of UA teachers matched to a comparison group of students based on student, teacher, and school characteristics. Our matching process relied on nearest neighbor with replacement and entropy balancing to create a comparison group with the same observable characteristics as the treatment group. These techniques enable us to use observational data to obtain "balance on covariates" between treatment and comparison groups. Nearest neighbor matching with replacement matches control individuals to the treated group and discards controls not selected. We allow control students to be matched to up to five treated students to ensure the best possible match and prevent matching order from affecting the match quality. We limit the number of matches to five because more unique control observations help to decrease the variance between observations. Entropy balancing reweights the observations to further balance the covariates and drops the observations furthest away in the covariate distribution. Students are matched exactly on categorical variables: year, eligibility for free and reduced lunch, race/ethnicity, female, English language learner, disability status, and home language other than English. The nearest neighbor match uses seventh grade performance on the ELA and mathematics exam (standardized by grade-year to mean zero standard deviation one); teacher characteristics: a set of indicators for years teaching at the NYCDOE (less than one year, 2-3 years, 4-5 years, 6-10 years, and 10 or more year), and two sets of indicator variables for license subject and assignment subject; and school characteristics: a set of indicator variables for the borough where school is located, total enrollment, percent of students who are Black, Latino, Asian, White, and multiracial, and percent of students who are economically disadvantaged. We use the Stata command *kmatch* (Jann, 2017b).

We identified 363,033 students (27.7% UA, 72.3% non-UA) who took the eighth grade ILS between 2012-13 and 2018-19 and could be matched to their eighth-grade science teacher. After matching, our analytical sample includes 232,399 students (44.5% UA, 55.5% non-UA) who took the eighth grade ILS exam from school years 2013-2019.

Figure 1 shows that, after matching, the treatment and comparison groups are balanced at baseline. The graph shows the standard mean difference and the variance ratio between the treatment and control groups for each variable used in matching; the blue dots reflect the raw data while the red dots reflect the matched sample. The matched sample shows a standard mean difference of approximately zero and a variance ratio close to one, indicating a good match (Jann, 2017a).

Table 1 presents information on the number of students in our matched sample, by year, whether they are in a UA school, and whether they are taught by a UA teacher. Across all years, approximately 66.9% of all students who took the ILS exam are enrolled at a UA school (Column 3), and of these, two-thirds are taught by a UA teacher (Column 5). Table 2 presents the number of schools in our sample, again by year and UA status. Our analysis includes a total of 534 unique schools that enroll eighth grade students over the seven-year period from 2013 through 2019. Of these, slightly more than three-quarters are identified as a UA school (a school with at least one active UA teacher in that year). Table 2 also presents the number of UA teachers in UA schools (Column 3). Of the 1,750 UA teachers (across all years), we were able to match 38.9% to students who have eighth grade ILS scores (Column 4). Most of the teachers who could not be matched teach science in grades other than eighth (e.g. sixth or seventh), teach other subjects, or teach specific populations (in particular, special education and bilingual education). Finally, of the UA teachers matched to students with scores on the ILS exam, almost

all (99.4%) are included in our matched sample used for analysis (Table 2 Column 5).

Table 3 presents descriptive statistics on the student sample by UA status. Columns 1 and 2 compare students attending schools that are and are not UA for the total sample. Columns 3 and 4 compare students at UA and non-UA schools in our analytic (matched) sample and Columns 5 and 6 take the subset of UA schools and compare students who are taught or not taught by a UA teacher. Taken as a whole, the demographic landscape of our sample of eighth grade ILS exam takers is typical of that of NYC public schools. Roughly three-quarters of students are eligible for free/reduced price lunch, there are slightly fewer females compared to males, and roughly 41% of students are Latino and 25% are Black. There are, however, important differences between students in the same school who do and do not have a UA teacher (Columns 5 and 6). Students taught by a UA teacher are more likely to be Black and less likely to be Asian or an English language learner. They also have lower average scores on the seventh grade ELA exam. Finally, they are taught by teachers with slightly fewer average years of teaching experience (8.7 versus 10.2).

Analytic Approach

Participation in the UA program is not random and depends on both observable and unobservable student, teacher, and school-level characteristics. To participate in UA, principals must first apply for their school to participate. Once the school is accepted into the UA program, individual teachers decide whether to participate. All sixth, seventh, or eighth grade teachers are eligible, regardless of the grade configuration of the school. Participating schools vary widely in student composition in terms of performance, poverty, and other sociodemographic characteristics, that is, both high and low performing schools, and high and low poverty schools, participate in UA. While principal buy-in starts the process, individual teachers have their own

reasons for choosing whether or not to participate. Anecdotal evidence from UA staff suggests that the number of years teaching and proximity of the school to a participating UA institution can influence participation. Additionally, because many of the PD sessions occur on the weekends, teachers may have personal obligations that prevent them from participating. Since teacher quality and experience are also strong predictors of student achievement (Ladd & Sorensen, 2017; Harris & Sass, 2011) this may introduce bias when estimating the relationship between UA exposure and student achievement.

Our matched sample attempts to account for bias introduced by selection into UA at the school and teacher level that is correlated with observable school and teacher characteristics, we also directly control for these characteristics in our regression analyses. In addition, to account for differential selection into UA across schools, we use a school fixed-effect and exploit variation in program participation across students, within schools, over time.

Specifically, we estimate model (1) to explore the relationship between having a UA teacher and eighth grade science achievement:

$$Y_{ijst} = \beta_0 + \delta_1 UA_{ijts} + X'_{ijst}\,\beta + \alpha_s + \tau_t + \varepsilon_{ijst} \qquad (1)$$

In this model, *Y* is the outcome of interest (ILS z-score score or an indicator for proficiency) for student *i* taught by teacher *j* in school *s* in year *t. UA*, our key variable of interest, is an indicator equal to 1 if a student has a UA teacher in the eighth grade and 0 if not. *X* is a vector of student and teacher characteristics that can influence UA participation and academic performance. Student characteristics include free/reduced price lunch eligibility, gender, race/ethnicity, SWD and ELA status, and scores on seventh grade statewide standardized math and ELA exams. Teacher characteristics include indicators for years of teaching experience and an indicator for whether they have a license in science. This vector of controls also includes a control for school

size (log enrollment). Lastly, α_s and τ_t are school and year fixed effects, respectively, and ε is the error term with the usual properties. Standard errors are clustered by teacher to account for correlation in outcomes across students within the same class. In models when the outcome is a binary indicator for proficiency (instead of the continuous science z-score), we estimate a linear probability model.[2]

Prior performance is a key predictor of future performance and may also correlate with student assignment to a UA teacher. While the NYS science exam is only given once in middle school (in eighth grade) wee do, however, observe and use seventh grade ELA and math exam scores as proxies for prior performance. Furthermore, the school fixed effect allows us to compare the performance of students with and without a UA teacher within the same school, and thus accounts for differences across schools in their likelihood to participate in the program and influence student performance, such as average academic performance levels, location, or principal leadership, that are relatively constant over time.

Results

Table 4 presents results from estimating Equation 1 on eighth grade ILS z-scores (Columns 1 and 2) and proficiency (Columns 3 and 4). We are interested in whether students who have a UA teacher in eighth grade outperform those who do not. For each outcome, we first estimate Equation 1 without school fixed effects. We add school effects, which is our preferred specification, in Columns 2 and 4 to capture differences in performance between students with and without UA teachers in the same school. While we see no statistically significant findings for z-score for UA students across schools (Column 1), we also see that UA students are 1.5 percentage points (pp) less likely to score as proficient on the exam compared to non-UA

[2] We estimate a linear probability model, as opposed to logit or probit, for ease of interpretation and because it has better consistency properties with fixed effects (Wooldridge, 2010).

students (Column 3). However, the estimates are notably different once we add school fixed effects. UA students perform 0.018 sd higher compared to non-UA students in the same school. This magnitude is relatively small and represents moving a UA student from scoring a 64.4 on the exam to 64.7. Unsurprisingly, then, we see no differences in the likelihood of students scoring proficient between UA and non-UA students in the same school (Column 4).

The change in the estimates when we include school fixed effects suggests that there is significant school-level selection that negatively biases the estimates of UA's impact in models that do not account for this selection. Put differently, there are unobserved features of schools that select into UA that are negatively correlated with student performance in science. This could, for example, be due to schools where students perform poorly on science selecting into UA in order to improve their students' science outcomes. As previously discussed, there are potentially many unobserved reasons for school selection into the UA program that also impact student outcomes, but if this selection bias is time-invariant, our school fixed effect eliminates this bias in the estimate.

In Table 5 we disaggregate the results presented in Table 4 (for the models with school fixed effects) by student characteristics. Columns 1 through 12 reveal the relationship between exposure to a UA teacher and students' science achievement is positive for many subgroups of students.[3] Students who are poor, Asian, White, male, SWDs, and ELLs all score between 0.02 and 0.05 sd higher than similar students in the same school without a UA teacher. SWDs taught by UA teachers are 1.5 pp more likely to be proficient on the exam, and ELLs taught by UA teachers are 1.9 pp more likely to be proficient on the exam, compared to similar students in their

[3] When conducting our sub-group analyses (e.g. estimating the effect for students of particular race/ethnicity), we re-matched students in our sample to students in the same subgroup to create a panel with the correct weighting to conduct the subgroup analysis.

school without a UA teacher. While we find Black students with a UA teacher are 2.6 pp less likely to be proficient in science compared to Black students in the same school without a UA teacher, this is the only subgroup for which we find a negative effect.

The analysis thus far suggests that UA exposure is associated with higher eighth grade science achievement, with larger impacts for SWDs and ELLs. We next estimate the impact of having a UA teacher on ELA or math. Given that UA focuses specifically on science, we would not expect significant impacts in these other subjects (that is, this can be viewed as a placebo test). Using the students in our matched sample to estimate impacts on ELA and math scores, we find no statistically significant differences for ELA and mathematics z-scores or proficiency (Table 6). This suggests that the estimates of the impact on student's science outcomes identify the true impact of the UA program and are not confounded by omitted variables.

Discussion and Conclusion

This study revisits prior work that estimated the impact of attending a UA school on students' science achievement. We take advantage of newly available teacher-student linkage and course data to examine whether having a UA teacher is associated with higher performance on the eighth-grade science exam. Overall, we find no differences in achievement between students with and without a UA teacher across schools, even when controlling for teacher level characteristics that may be correlated with both UA program participation and student outcomes (teaching experience and licensing). However, given there is also *school-level* selection into the UA program that can bias results, we estimate models with school fixed effects and do find within school differences: students in a school who have a UA teacher in eighth grade score 0.02 sd higher compared to other students without a UA teacher in the same school. Although this is lower than the impacts found in some prior studies of science professional development on

student achievement, most of these assessments do not use a state standardized science exam to evaluate effects. Our subgroup analyses also suggest that UA can be especially helpful for SWD and ELL students. Since some components of the UA program are implemented at the school level, and program staff often consider it a whole-school intervention (e.g. it requires principal buy-in, engages parent coordinators, and provides field trips for all students—not just those with a UA teacher), these differences between students within the same school suggest a particular importance of the PD component and the within school participation of teachers.

The UA program is a unique partnership made possible through an ongoing collaboration between eight science-rich cultural organizations and the NYCDOE. As a long-standing collaborative program, evidence from UA has implications not only for improving science teaching but also more generally for creating stronger partnerships between school districts and external institutions. Moreover, the analyses highlight the importance of revisiting previous studies when new information, in this case more nuanced data to better identify treatment, becomes available.

There are important limitations of our analysis that mean we cannot confirm whether estimated relationships are causal. We cannot, for example, identify other factors such as teacher motivation or location preferences that may also influence teaching quality and the decision to participate in UA. It is also worth noting that this study does not identify specific mechanisms through which UA may improve science achievement. The benefits of UA can flow through many channels, such as higher engagement and motivation among students or a more collaborative school environment for educators. Future work can empirically investigate these potential mechanisms to provide guidance on policy and practice for districts that may wish to implement similar programs. Additional evidence on the benefits of such collaborations, for both

formal and informal stakeholders, and on specific mechanisms of change can help institutionalize productive partnerships between schools and external educational institutions.

References

Allen, L. B. & Crowley, K. J. (2014). How museum educators change: Changing notions of learning through changing practice. *Science Education, 98*(1), 84-105. https://doi.org/10.1002/sce.21093

Ash, D. B. & Lombana, J. (2012). Methodologies for reflective practice and museum educator research: The role of noticing and responding. In D. B. Ash, J. Rahm, & L. M. Melber (Eds.), *Putting theory into practice: Tools for research in informal settings*. (pp. 29-52). SensePublishers. https://doi.org/10.1007/978-94-6091-964-0_4

Bevan, B. Dillon, J., Hein, G.E., Macdonald, M., Michalchik, V., Miller, D., Root, D., Rudder, L., Xanthoudaki, M., & Yoon, S. (2010). Making Science Matter: Collaborations Between Informal Science Education Organizations and Schools. A CAISE Inquiry Group Report. *Center for Advancement of Informal Science Education (CAISE).* https://www.informalscience.org/making-science-matter-collaborations-between-informal-science-education-organizations-and-schools

Cheung, A., Slavin, R.E., Kim, E., Lake, C. (2017). Effective secondary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching, 54*(1), 58–81. https://doi.org/10.1002/tea.21338

Greene, J. P., Kisida, B., & Bowen, D. H. (2014). The educational value of field trips. *Education Next,* 14(1), 78-86. https://www.educationnext.org/the-educational-value-of-field-trips/

Harris, C. J., Feng, M., Murphy, R., & Rutstein, D. W. (2022). Curriculum materials designed for the Next Generation Science Standards show promise: Initial results from a randomized controlled trial in middle schools. *WestEd.* https://www.wested.org/resources/curriculum-materials-for-ngss/

Harris, D. N. & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7), 798-812. https://doi.org/10.1016/j.jpubeco.2010.11.009

Jann, B. (2017a). "kmatch: Kernel matching with automatic bandwidth selection," United Kingdom Stata Users' Group Meetings 2017 11, Stata Users Group. https://ideas.repec.org/p/boc/usug17/11.html

Jann, B. (2017b). "KMATCH: Stata module for multivariate-distance and propensity-score matching, including entropy balancing, inverse probability weighting, (coarsened) exact matching, and regression adjustment." Statistical Software Components S458346, Boston College Department of Economics, revised 19 Sep 2020. https://ideas.repec.org/c/boc/bocode/s458346.html

Krajcik, J., Schneider, B., Miller, E., Chen, I.-C., Bradford, L., Baker, Q., Bartz, K., Miller, C., Li, T., Codere, S., & Peek-Brown, D. (2022). Assessing the Effect of Project-Based Learning on Science Learning in Elementary Schools. *American Educational Research Journal*, *60*(1), 70-102. https://doi.org/10.3102/00028312221129247

Lacoe, J., Painter, G. D., & Williams, D. (2020). Museums as Classrooms: The Academic and Behavioral Impacts of "School in the Park." *AERA Open*, *6*(3). https://doi.org/10.1177/2332858420940309 .

Ladd, H. F. & Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy*, *12*(2), 241-279. https://doi.org/10.1162/EDFP_a_00194

Loucks-Horsley, S., Stiles, K. E., Mundry, S. E., Love, N., & Hewson, P. W. (2010). *Designing professional development for teachers of science and mathematics*. Corwin Press.

https://doi.org/10.4135/9781452219103

Loucks-Horsely, S. & Matsumoto, C. (1999). Research on Professional Development for Teachers of Mathematics and Science: The State of the Scene. *School Science and Mathematics*, *99*(5), 258-271. https://doi.org/10.1111/j.1949-8594.1999.tb17484.x

National Research Council. (2009). *Learning Science in Informal Environments: People, Places, and Pursuits.* The National Academies Press. https://doi.org/10.17226/12190.

Phillips, M., Finkelstein, D., & Wever-Frerichs, S. (2007). School site to museum floor: How informal science institutions work with schools. *International journal of science education*, *29*(12), 1489–1507. https://doi.org/10.1080/09500690701494084

Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Klager, C., Bradford, L., Chen, I.-C., Baker, Q., Touitou, I., Peek-Brown, D., Dezendorf, R. M., Maestrales, S., & Bartz, K. (2022). Improving Science Achievement—Is It Possible? Evaluating the Efficacy of a High School Chemistry and Physics Project-Based Learning Intervention. *Educational Researcher*, *51*(2), 109–121. https://doi.org/10.3102/0013189X211067742

Seraphin, K. D., Harrison, G. M., Philippoff, J., Brandon, P.R., Nguyen, T.T.T., Lawton, B.E., & Valin, L.M. (2017). Teaching Aquatic Science as Inquiry Through Professional Development: Teacher characteristics and student outcomes. *Journal of Research in Science Teaching*, *54*(9), 1219-145. https://doi.org/10.1002/tea.21403

Weinstein, M., Whitesell, E. 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. https://doi.org/10.1177/0193841X14553299

Whitesell, E.R. (2016). A day at the museum: The impact of field trips on middle school science achievement. *Journal of Research in Science Teaching*, *53*(7), 1036-1054.

https://doi.org/10.1002/tea.21322

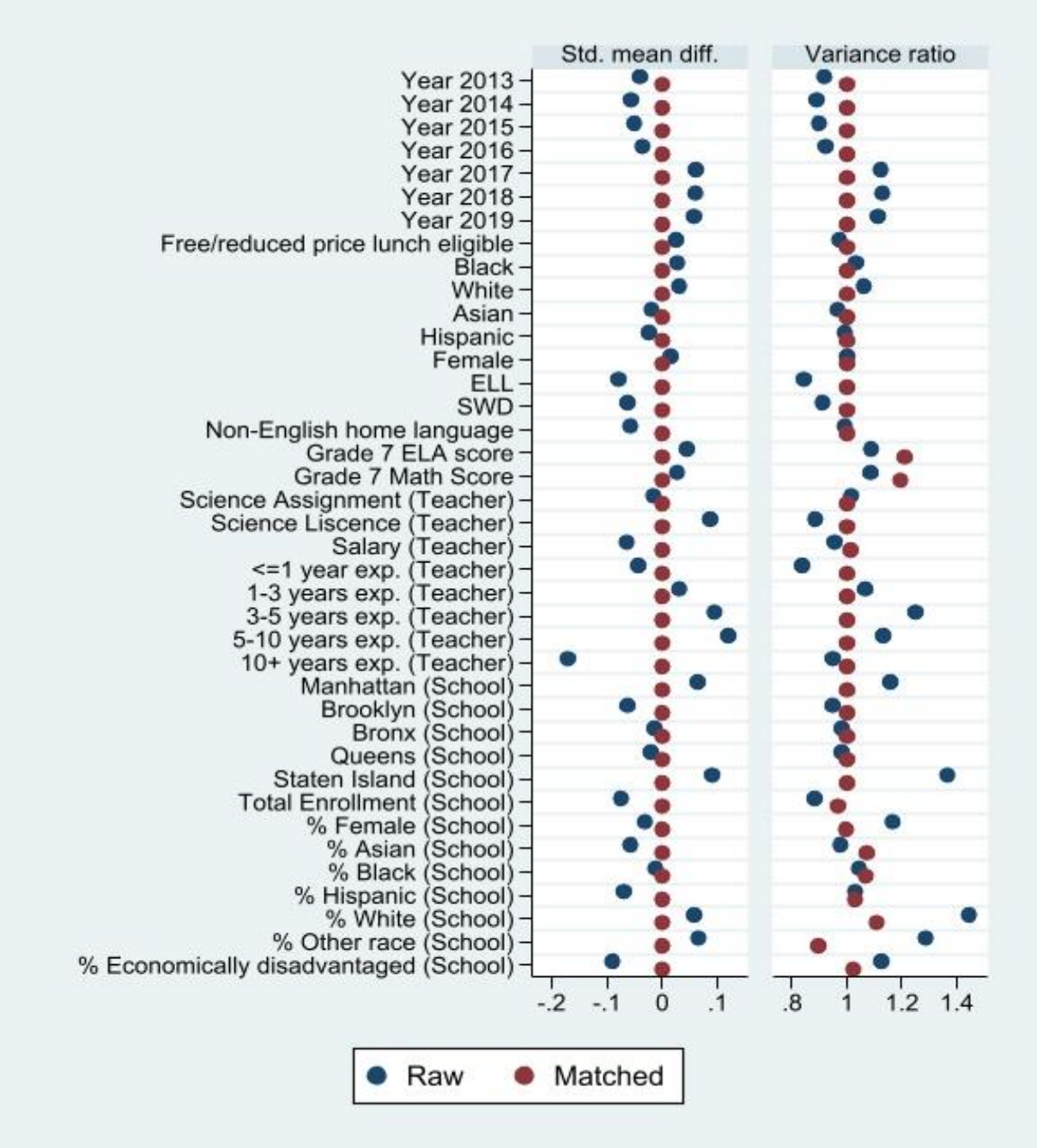
Woolridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data.* The MIT Press. https://www.jstor.org/stable/j.ctt5hhcfr

Yerrick, M., & Beatty-Adler, D. (2017). Addressing Equity and Diversity with Teachers Though Informal Science Institutions and Teacher Professional Development. *Journal of Science Teacher Education, 22*(3), 229-253. https://doi.org/10.1007/s10972-011-9226-3

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement (Issues & Answers Report, REL 2007–No. 033). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/edlabs

Zoblotsky, T., Bertz, C., Gallagher, B., & Alberg, M. (2017). The LASER Model: A Systematic and Sustainable Approach for Achieving High Standards in Science Education. SSEC i3 Validation Final Report of Confirmatory and Exploratory Analyses [Updated]. *Center for Research in Educational Policy (CREP)*. https://files.eric.ed.gov/fulltext/ED617222.pdf

Figure 1. Quality of Matching for Analytic Sample



Note. The graph shows the standard mean difference and the variance ratio between the treatment and control groups for each student-, teacher-, and school-level variable used in matching; the blue dots reflect the raw data while the red dots reflect the matched sample. The matched sample shows a standard mean difference of approximately zero and a variance ratio close to one, indicating a good match (Jann, 2017a).

Table 1. Students in the analytic sample, by year and UA school

Year	Total Students in Sample	Students Enrolled in UA Schools		Students Taught by a UA Teacher in UA School	
	(1)	(2)	(3)	(4)	(5)
	N	N	% of Total	N	% of Students in UA Schools
2013	31,917	21,775	*68.2*	13,399	*61.5*
2014	33,336	22,145	*66.4*	13,697	*61.9*
2015	32,214	19,246	*59.7*	13,328	*69.3*
2016	29,706	16,699	*56.2*	12,513	*74.9*
2017	35,608	24,436	*68.6*	17,118	*70.1*
2018	33,420	24,254	*72.6*	16,071	*66.3*
2019	36,198	26,879	*74.3*	17,290	*64.3*
Total:	232,399	155,434	*66.9*	103,416	*66.5*

Note. Sample includes eighth grade students who took the New York State Intermediate Level Science (ILS) exam and who could be matched to their science teacher from the 2013-2019 school years. Special education only schools, charter schools, schools with less than 10 tested students, and teachers with less than 10 tested students are excluded.

Table 2. Schools and teachers in the analytic sample, by year

	(1)	(2)		(3)	(4)		(5)	
	Total Schools	**UA schools**		**UA Teachers in Current Year**	**UA Teachers Matched to Student Test Scores**		**UA Teachers Matched to Student Test Scores in Analysis**	
	N	**N**	**%**	**N**	**N**	**%**	**N**	**%**
2013	439	198	*45.1*	356	176	*49.5*	173	*98.3*
2014	444	214	*48.2*	501	204	*40.7*	201	*98.5*
2015	443	154	*34.8*	617	209	*33.9*	206	*98.6*
2016	426	145	*34.0*	508	179	*35.2*	179	*100*
2017	430	257	*59.8*	765	281	*36.7*	281	*100*
2018	412	261	*63.3*	781	270	*34.6*	265	*98.1*
2019	384	275	*71.6*	845	290	*34.3*	290	*100*
Total (unique)	534	418	*78.3*	1,750	681	*38.9*	677	*99.4*

Note. Column 1 is the total number of schools in the analytic sample (special education only schools, charter schools, and schools with less than 10 tested students are excluded). Column 2 is the total number of UA schools and the corresponding percentage as a share of Column 1. Column 3 is the number of all UA teachers in NYC traditional public schools each year who could be matched to NYC teacher data. Column 4 is the number of UA teachers we are able to match to students with eighth grade test scores (using the student-teacher linkage data) and the corresponding percentage as a share of Column 3. The majority of unmatched teachers have students in grades that do not take the ILS exam. Column 5 is the number of UA teachers in our matched analytic sample and the corresponding percentage as a share of column 4.

Table 3. Characteristics of students in analytic sample: NYC eighth grade students in traditional public schools from 2013-2019

	All Students		**Analytic (Matched) Sample**		**Students in UA Schools Only (Analytic Sample)**	
	UA	Not UA	UA	Not UA	UA teacher	No UA teacher
	(1)	(2)	(3)	(4)	(5)	(6)
Percent of students who are:						
Free/reduced price lunch eligible	75.3	76.7	75.4	74.3	75.4	74.2
Female	47.7	47.5	47.6	46.8	47.6	46.2
Latino	41.4	43.0	41.3	42.6	41.3	40.1
Black	25.4	25.1	25.4	24.2	25.4	20.9
White	15.4	13.7	15.5	14.4	15.5	15.3
Asian	16.1	16.6	16.2	16.9	16.2	21.1
Multiracial or other	1.3	1.0	1.0	1.0	1.0	1.4
Student with disabilities	19.0	17.5	18.9	21.3	18.9	20.0
English language learner	12.0	14.1	11.9	14.6	11.9	16.1
Language other than English	44.9	48.6	44.9	47.8	44.9	51.4
Taught by UA teacher	28.9	0.0	44.5	0.0	66.5	33.5
Outcomes and teaching experience:						
Average z-score ELA, Grade 7	-0.05	-0.05	-0.05	-0.09	-0.05	-0.10
Average z-score math, Grade 7	-0.07	-0.05	-0.07	-0.09	-0.07	-0.07
Proficient on science exam	52.8	53.9	52.9	52.6	52.9	53.5
Average Science z-score	0.01	0.02	0.01	-0.01	0.01	0.01
Average ELA z-score, Grade 8	-0.4	-0.03	-0.03	-0.07	-0.03	-0.07
Average Math z-score, Grade 8	0.01	0.02	0.01	-0.02	0.01	0.20
Years of teaching experience	8.6	9.3	8.7	9.7	8.7	10.2
Number of observations	104,851	258,182	103,416	128,983	103,416	52,018
Number of schools	425	482	418	469	418	0.0
Number of teachers	681	2288	677	2160	677	1241

Note. Sample includes eighth grade students who took the New York State Intermediate Level Science (ILS) exam and who could be matched to their science teacher in years 2013-2019. Special education only schools, charter schools, schools with less than 10 tested students and teachers with less than 10 tested students are excluded. The z-score is a measure of relative performance on the ILS exam standardized across students within a grade and year to have a mean of 0 and standard deviation of 1. All characteristics in Columns 1 and 2 are statistically different from one another at the 5% significance level except for the percent of students who are female. All characteristics in Columns 3 and 4 are statistically different from one another.

Table 4. Results: Having a UA teacher in eighth grade on science achievement, 2013-19

	Z-Score		Proficiency	
	(1)	(2)	(3)	(4)
Current UA Teacher	-0.017	0.018*	-0.015**	0.003
	(0.015)	(0.011)	(0.007)	(0.005)
Constant	-0.479***	0.394**	0.372***	0.807***
	(0.081)	(0.199)	(0.038)	(0.090)
Year Effects	Y	Y	Y	Y
Student Characteristics	Y	Y	Y	Y
Teacher Characteristics	Y	Y	Y	Y
School Fixed Effects	N	Y	N	Y
N	232,399	232,399	232,399	232,399
adj. R^2	0.562	0.609	0.405	0.442

Robust standard errors clustered by teacher in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note. Sample includes eighth grade students who took the New York State (NYS) Intermediate Level Science (ILS) Exam and could be matched to their science teacher in years 2013-2019. Special education only schools, charter schools, schools with less than 10 tested students, and teachers with less than 10 tested students are excluded from the sample. The outcome z-score is a measure of relative performance on the ILS exam standardized across students within grade and year to have a mean of 0 and standard deviation of 1. In NYS, the outcome proficient means scoring in performance levels 3 or 4. Columns 3 and 4 are linear probability models. Student characteristics include student's gender, race/ethnicity, participation in special education, English language learner, free/reduced price lunch eligibility, and seventh grade scores on NYS standardized math and English exams. Teacher characteristics include teacher's teaching experience, licensing, and an indicator for whether they taught science. All models also adjust for school size (log enrollment) and include indicators for student and teacher-level characteristics that were missing and replaced (all missing indicator variables were set to zero).

Table 5. Regression results by student characteristics, having a UA teacher in eighth grade and ILS achievement.

	(1) Poor	(2) Not Poor	(3) Black	(4) Latino	(5) Asian	(6) White	(7) Female	(8) Male	(9) SWD	(10) Not SWD	(11) ELL	(12) Not ELL
Panel A: Z-score												
Have UA teacher	0.019*	0.015	-0.022	0.016	0.028**	0.047**	0.009	0.023*	0.044***	0.004	0.041**	0.013
	(0.012)	(0.014)	(0.020)	(0.014)	(0.013)	(0.023)	(0.012)	(0.012)	(0.017)	(0.011)	(0.020)	(0.011)
N	173443	58947	54790	97369	38132	34956	110188	122361	47324	189656	30602	202413
Adjusted R^2	0.583	0.640	0.497	0.533	0.649	0.618	0.637	0.592	0.442	0.610	0.311	0.614
Panel B: Proficient												
Have UA teacher	0.005	-0.000	-0.026**	0.007	-0.005	0.016	-0.003	0.007	0.015*	-0.003	0.019**	-0.001
	(0.005)	(0.007)	(0.012)	(0.007)	(0.006)	(0.011)	(0.006)	(0.005)	(0.009)	(0.005)	(0.008)	(0.005)
N	173443	58947	54790	97369	38132	34956	110188	122361	47324	189656	30602	202413
Adjusted R^2	0.421	0.453	0.359	0.384	0.427	0.402	0.453	0.436	0.299	0.610	0.229	0.429
Year effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Student characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Teacher characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
School Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Robust standard errors clustered by teacher in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note. Sample includes eighth grade students who took the New York State (NYS) Intermediate Level Science (ILS) Exam and could be matched to their science teacher in years 2013-2019. Special education only schools, charter schools, schools with less than 10 tested students and teachers with less than 10 tested students are excluded from the analysis. The outcome z-score is a measure of relative performance on the ILS exam standardized across students within a grade and year to have a mean of 0 and standard deviation of 1. In NYS, the outcome proficient means scoring in performance levels 3 or 4. Each column (in each panel) is a separate regression by subgroup. Student characteristics include student's gender, race/ethnicity, participation in special education, English language learner, free/reduced price lunch eligibility, and seventh grade scores on NYS standardized math and English exams. Teacher characteristics include teacher's teaching experience, licensing, and an indicator for whether they taught science. All models also adjust for school size (log enrollment) and include indicators for student and teacher-level characteristics that were missing and replaced (all missing indicator variables were set to zero).

Table 6. Regression results, having a UA teacher in 8th grade and achievement on ELA and Math, 2013-19

	ELA		Math	
A. Z-Scores				
Current UA Teacher	-0.012	-0.005	0.017	0.016
	(0.008)	(0.009)	(0.021)	(0.018)
Constant	-0.063	0.795***	-0.404***	0.803***
	(0.051)	(0.159)	(0.118)	(0.297)
N	204522	204522	190841	190841
adj. R^2	0.610	0.622	0.321	0.406
B. Proficiency				
Current UA Teacher	-0.006	0.000	0.007	0.004
	(0.004)	(0.005)	(0.010)	(0.008)
Constant	0.228***	0.352***	0.162***	0.308**
	(0.027)	(0.085)	(0.053)	(0.128)
N	204522	204522	190841	190841
adj. R^2	0.414	0.423	0.226	0.293
Year Effects	Y	Y	Y	Y
Student Characteristics	Y	Y	Y	Y
Teacher Characteristics	Y	Y	Y	Y
School Fixed Effects	N	Y	N	Y

Robust standard errors clustered by teacher in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note. Sample includes eighth grade students who took the New York State (NYS) Intermediate Level Science (ILS) Exam and could be matched to their science teacher for the years 2013-2019. Excluded from the analysis are special education only schools, charter schools, schools with less than 10 students, and teachers with less than 10 students and in UA program for only one year. The outcome z-score is a measure of relative performance on the ILS exam standardized across students within a grade and year to have a mean of 0 and standard deviation of 1. In NYS, the outcome proficient means scoring in performance levels 3 or 4. Columns 3 and 4 are linear probability models. Student characteristics include student's gender, race/ethnicity, participation in special education, English language learner, free/reduced price lunch eligibility, and seventh grade scores on NYS standardized math and English exams. Teacher characteristics include teacher's teaching experience, licensing, , and an indicator for whether they taught science. All models also adjust for school size (log enrollment) and include indicators for student and teacher-level characteristics that were missing and replaced (all missing indicator variables were set to zero).