September 15, 2016

Evaluation of a Year-Long Hands-on Elementary Science Program

P.S. SCIENCE AND STUDENT ACHIEVEMENT

Greg Srolestar, MPP
GSROLESTAR@GMAIL.COM |

Table of Contents

Executive Summary

The P.S. Science program provides hands-on, inquiry-based instruction directly to elementary science students and provides concurrent professional support to their teachers over the course of a school year. The program seeks to provide low-income students with high-quality science education, consistent with Next Generation Science Standards. Meanwhile, teachers are provided with in- and out-of-classroom support designed to improve their instruction skills. This evaluation used a straightforward quantitative analysis to examine changes across the most important domain: student learning.

At the end of the 2015/16 school year, students in four 3rd grade P.S. Science classrooms at a Lawndale elementary school participated in a scientific vocabulary assessment. The demonstration of improved vocabulary is indicative of improved capacity for comprehension of scientific concepts. Three 3rd grade classrooms from a non-P.S. Science school took the same exam under the same conditions. Scores were compared between the two groups.

Students who participated in P.S. Science scored an average of 1.23 points higher on a grade-appropriate scientific vocabulary assessment than a control group of non-P.S. Science students. Converting this into standardized units, we see an effect size of +0.49 standard deviations, comparable to double the average effect of decreasing class size by nine students. The effect was statistically significant at the .01 level. Statistically controlling for student gender and student home language did not impact the results. The effect size is substantial and shows that P.S. Science students finished the year with stronger vocabularies than student in the control group. Further study is needed to more precisely determine the magnitude of the effect of P.S. Science on student vocabulary.

Background

As noted in previous evaluations of the P.S. Science program, "standards documents and national organizations continue to advocate for quality, inquiry-based science experiences at the elementary level in order to better prepare students for science learning at the higher grades and to support their development into scientifically literate adults (National Research Council, 2011; National Research Council 1996, National Science Teacher Association, 2002)" (Melber, 2012). These standards are codified in the Next Generation Science Standards, put out by the National Research Council in 2011, adopted by California's State Board of Education in 2013 (California Department of Education, 2016).

The purpose of these standards, as articulated by the Next Generation Science Standards consortium, is to prepare students for careers in a modern workforce, improve American scientific achievement relative to other nations, and build a more scientifically and mathematically literate citizenry (Next Generation Science Standards, 2016).

Consistent with the standards and best practices described above, the P.S. Science program provides early-grade students in several Los Angeles County low-income schools with inquiry-based science instruction. Students receive a year-long, hands-on science curriculum provided by P.S. Science content and teaching experts, known as instructors. Each week students receive a 90 minute science session that includes instruction in scientific inquiry along with, "the opportunity to engage in an aspect of active investigation or hands-on exploration using objects and specimens" (Melber, 2012).

Along with direct instruction for the students, the program provides classroom teachers with additional support (e.g. coaching, modeling, assistance extending lesson content to other discliplines, lesson materials) to improve their comfort and ability in teaching science. The classroom teacher is present throughout the session and participates as is appropriate; if the classroom is temporarily split for smaller group instruction, the teacher will lead one of those two groups. Teachers also receive out-of-classroom support as needed. Ensuring teachers improve their comfort with hands-on, inquiry-based science instruction is an essential component of the program. Previous research has found that elementary school teachers are less prepared to teach science than English or Math (California Council on Science and Technology, 2010).

Evaluation Objective

Previous qualitative evaluations have surveyed and interviewed classroom teachers, P.S. Science instructors, and students' parents and found the program meets its programmatic objectives. Families felt their children enjoyed and learned science, teachers were initially challenged by the content but grew more comfortable with it over time, and instructors felt that teachers were, "enthusiastic and capable," (Melber, 2012; Melber, 2013).

The objective of this evaluation is to add to the qualitative story by providing quantitative evidence of the impact of the P.S. Science program on student learning.

Methodology and Data Collection

To measure the impact of P.S. Science, we employed a standardized, grade-appropriate measure of student vocabulary knowledge from "Seeds of Science, Roots of Reading", created and psychometrically tested by the Lawrence Hall of Science and the Graduate School of Education at the University of California, Berkeley. Vocabulary is widely recognized as highly correlated with reading comprehension, though the causal chain is complex (Nagy, 2006). Unfamiliar scientific vocabulary may prevent students

from learning science, especially at older grades as the vocabulary load of scientific instruction grows. (Carnine & Carnine, 2004). The demonstration of improved vocabulary is indicative of improved capacity for comprehension of scientific concepts.

To determine if students in the P.S. Science program acquire improved scientific vocabulary, we sought to compare students receiving P.S. Science (the treatment group) to similar students not receiving the P.S. Science program (the control group). Third grade students at a Lawndale elementary school that received the P.S. Science program and at a matched Compton elementary school took a standardized 12 question vocabulary assessment and a companion short answer question.[1] At the end of the assessment (to reduce bias from stereotype threat), students provided information on their gender and languages spoken at home, allowing for subgroup analysis. No identifying student data such as names or student id numbers was collected.

The assessment was proctored by a P.S. Science staff person in the control group and the classroom teacher in the P.S. Science group. Students did not receive a lesson on the specific vocabulary in assessment, either before or concurrent with the assessment. Neither the teacher nor the P.S. Science instructor were aware of the questions until the week of the assessment and did not prepare students in advance. Both proctors followed a specific assessment protocol, including reading the questions aloud and instructing students to answer questions independently.

Assessments were collected by the proctor and returned to P.S. Science to be graded and tabulated. Mean student scores from the treatment and control groups were compared using a two-tailed t-test. The t-test provides us with some level of confidence that our two groups are truly different and that the difference between the two scores is not simply a function of chance. More technically, we test the null hypothesis that there is no difference in scientific vocabulary knowledge between the two groups. Further robustness testing was performed to strengthen our confidence in the result.

In the next section, we examine the similarities and differences between the students in the treatment and control groups. The more similar the groups, the stronger our analysis.

School and Student Data Review

To determine whether the students are similar, we first examine schoolwide data. Public data on each school from 2015/16 shows broadly similar racial and ethnic demographics with the control group having a higher percentage of children eligible for Free or Reduced Price Meals and who are English Learners.

[1] The Compton control group school is a likely candidate for P.S. Science in the upcoming year and agreed to allow students to participate in the assessment to assist with the evaluation.

Table 1 School-Wide Demographics

	Category	William Green (treatment)	Dickison (control)	Difference
Demographics	School Population	765	827	-62
	Black or African American	6%	11%	-5%
	American Indian or Alaska Native	0%	0%	0%
	Asian	1%	0%	1%
	Filipino	1%	0%	1%
	Hispanic or Latino	82%	88%	-5%
	Native Hawaiian or Pacific Islander	0%	0%	0%
	White	7%	1%	6%
	Two or More Races	2%	0%	2%
	Free or Reduced Price Meals	82%	91%	-9%
	English Learners	43%	55%	-11%

Data on 3rd grade students at both schools shows larger demographic differences. The treatment school is approximately 12% White and Asian students while the control group has almost no White or Asian students. Likewise the share of both English Language Learners and economically disadvantaged students is larger at the control group school. 2015/16 test scores show a larger percentage of treatment group students meeting or exceeding standards in both English and Math, with large differences in English scores.

Table 2 Third Grade Demographics and 15/16 Test Scores

	Category	William Green (treatment)	Dickison (control)	Difference
Demographics	Third Grade Population	116	131	-15
	Black or African American	8%	11%	-3%
	American Indian or Alaska Native	0%	0%	0%
	Asian	3%	0%	3%
	Filipino	0%	0%	0%
	Hispanic or Latino	78%	89%	-10%
	Native Hawaiian or Pacific Islander	1%	0%	1%
	White	8%	0%	8%
	Two or More Races	2%	0%	2%
	Free or Reduced Price Meals	84%	95%	-11%
	English Learners	46%	61%	-15%
Tests	ELA CAASP % Meeting or Exceeding Standards	37%	20%	17%
	Math CAASP % Meeting or Exceeding Standards	34%	26%	8%

Next, we turn our attention toward the specific group of students who took the assessment. For these students, we collected data on student gender and the percent speaking English at home. The data shows that slightly more children in the control group did not speak English at home and that the treatment group had a larger percentage of female students.

Table 3 Treatment/Control Classroom Differences

	Lawndale (treatment)	Compton (control)	Difference
% Female	53.5%	42.0%	11.5%
% No English at Home	27.4%	32.8%	-5.4%

In total, the data indicate that students in the control group are more likely to be from groups that struggle academically. This challenges our analysis in that it biases results upward, making it difficult to conclusively attribute improvement in student science vocabulary test scores to P.S. Science.

Results

In this section we present compelling evidence that P.S. Science students finish the year with stronger science vocabularies than the control group students. As Table 4 Summary Statisticsindicates, there were 70 students in the control group and 87 in the P.S Science group. *The raw mean scores for the P.S. Science group were 1.23 points higher than the for the control group, meaning that the tested P.S. Science group outperforms the tested control group.*

Table 4 Summary Statistics

	PS Science Students	Control Students
Number	87	70
Number of Classrooms	4	3
Percent Female	53.5%	42.0%
% No English at Home*	27.4%	32.8%
Mean Vocab Score	8.57	7.34
Std Vocab	2.49	2.52

*Refers to the percentage of students who report speaking *only* a language other than English at home.

We would also like to know the answer to the more general question, “Do students that receive P.S. Science know more scientific vocabulary than those who do not?” In other words, if we examined other groups of students that had taken P.S. Science and those who had not, would we find similar vocabulary test score differences? To answer this question we need to treat our groups to as a representative sample of P.S. Science students and a representative sample of non-P.S. Science students. Ideally the two groups would be identical except that one group would receive P.S. Science. While this is clearly not the case, both groups of students come frome Title 1 schools, are overwhelmingly African American and

Latino, students do not self-select into the program, and both schools have chosen to participate in P.S. Science.[2]

A two-tailed t-test allows us to generalize from our sample, while demonstrating that our finding is not a matter of chance. More technically, the t-test allows us to reject the null hypothesis at some confidence level that the scores of control students and P.S. Science students are equivalent. In order to employ a t-test, our groups should meet several conditions. First, we need to evaluate two groups along a continuous dependent variable, in this case vocabulary scores. Second, scores must be independently observed, meaning scores are not correlated between the two groups. Third, the variances of the two samples must be approximately equivalent. Fourth, scores must be distributed normally within population groups. Lastly, ideally both samples are randomly drawn from the population. Our samples meet each of the criteria except for the last. [3] For more information, see Technical Appendix.

Running a two-tailed t-test provides evidence that P.S. Science students outperform similar non-P.S. Science students. We can reject the null hypothesis with over 99% confidence (p value = 0.0026, t = 3.065).

A further important piece of information is the size of the effect. A standard approach for evaluating effect sizes is to standardize the differences between the means (Cohen's *d*). Putting the effect size into standardized units allows for comparisons across interventions.

The Cohen's *d* effect size in this instance is +0.49, nearly half a standard deviation. In education, an effect size of .5 is considered medium to large, depending on the intensiveness of the intervention. By way of comparison, Tennessee Class Size Experiment found that shrinking class from 22-26 students to 13-17 students produced an effect of 0.11 in reading and +0.22, considerably less than the effect found here (Bloom, Hill, Rebeck Black, & Lipsey, 2006). A more relevant comparison can be made to a 2007 elementary science vocabulary intervention which found an effect size of +0.24, significantly less than in P.S. Science (Hanley, Lake, Slavin, & Thurston, 2012). Several interventions that focused on professional development in inquiry-oriented science showed consistent positive effects on elementary science achievement of approximately +0.3 (Hanley, Lake, Slavin, & Thurston, 2012).

Robustness Test

One concern is that the difference in scores between the two groups reflects underlying demographic differences. The P.S. Science treatment group is more heavily female and, on our exam, girls score higher than boys on average (8.49 and 7.65, respectively). Likewise P.S. Science students are more likely to speak English in the home, with students who speak no English in the home scoring lower (8.13 and 7.83, respectively).

Multiple regression analysis allows us to "control for", or hold constant, the effects of gender and language and look just at the impact of P.S. Science. In other words, we are asking what is the effect of taking P.S. Science if we statistically adjust to remove the impacts of gender or home language? The below regression model only accounts for a small number of variables impacting student vocabulary;

[2] The control group is

[3] Some analysts prefer not to perform a t-test on groups that are not randomly sampled. Random sampling was not an option in this case. We present both the raw data and the t-test results so the reader may decide for herself, which data are most useful.

however, they are important variables.[4] We are able to see that controlling for gender and home language does not reduce the estimate of the impact of P.S. Science. Where the raw scores indicate a mean difference of 1.23 points, the regression model finds that holding gender and home language constant, participating in P.S. Science is associated with a mean score increase of 1.22 points. The p-value is very low, indicating statistical significance at higher than a 99% confidence level. The below regression illustrates that controlling for gender and home language has no meaningful impact on statistical significance or effect size. This is important because while we cannot control for all of the observed differences between the treatment and control groups, when we can control for differences, the effect remains strong. See Technical Appendix for coding details.

Table 5 Vocabulary Score Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	7.0646383	0.37737814	18.72032	3.419E-41
PS Student	1.2193389	0.40317204	3.024364	0.0029289
Gender	0.7038812	0.39935096	1.762563	0.0799974
Home Language	-0.0720231	0.42373958	-0.16997	0.8652612

Discussion

The finding that P.S. Science increases scientific vocabulary in third grade students by 1.2 points or Cohen's *d* of +0.49 is suggestive of a very impactful program, much larger than a comparable 12-week elementary science vocabulary intervention. In interpreting these results, it is important to note that a different methodological design including randomization or pre-test/post-test with control group may attenuate results. Additionally, this sample is not representative of all students nationwide. Different groups of students may respond differently to the P.S. Science program. Still, the large, positive results should be interpreted as suggestive evidence of a program that can successfully improve scientific vocabulary among 3rd graders.

There are reasons to believe P.S. Science produces a large positive effect though more evidence is needed. Several elements of the P.S. Science program should make us accepting of the large effects. First, professional development provided by content experts has been shown to significantly improve results in elementary science (Hanley, Lake, Slavin, & Thurston, 2012). Researchers have found that elementary teachers often lack scientific content knowledge, which impedes their ability to provide inquiry-based instruction (Luera, 2005). Second, the P.S. Science program is particularly intensive, providing direct instruction to students and hands-on support to teachers for the entire school year. Lastly, small programs, such as P.S. Science run by a passionate and knowledgeable team often outperform interventions at scale.

[4] Ideally we would have data on individual student race/ethnicity and free or reduced lunch status. We chose not to ask students about those domains for two reasons. First we worried 3rd graders would not accurately report their statuses. Second, we did not wish to introduce topics that can lead to complex or difficult conversations.

Should the P.S. Science program conduct future evaluations, there are interesting avenues for further study. First, a repeat study can demonstrate consistency of program impact, ideally with more similar treatment and control groups. A different methodology, perhaps pretest/posttest of treatment and control groups, can lend increased statistical rigor to the findings. Lastly, it would be interesting to follow teachers who have benefitted from P.S. Science but moved on to other schools. Would their future students who did not participate in the P.S. Science program benefit from their teacher's involvement in the program? Given the promising findings, considering to measure and monitor the results of P.S. Science is particularly valuable.

Technical Appendix

Test for normality

We should have a strong presumption of normality in student test scores. As seen in Figure 1, a histogram illustrates sufficient normality for the control group (though it is slightly Mesokurtic). The PS Science group does not obviously meet the standards for normality, with Figure 2 showing negative skew. There are, however, two mitigating factors. First, small samples are less likely to meet the condition of normality as a matter of chance. Second, ceiling effects appear to be skewing the distribution. If the assessment had more questions, it appears likely that there would be less clustering at the top of the distribution and a relatively normal curve would emerge. Importantly, these ceiling effects bias downward the results for the PS Science group, suggesting that an exam with more questions might show a larger effect of the PS Science program. The test for normality is a judgement call and given the below histograms and presumption of normality in students taking a normed exam, we believe this meets the normality test.

Figure 1

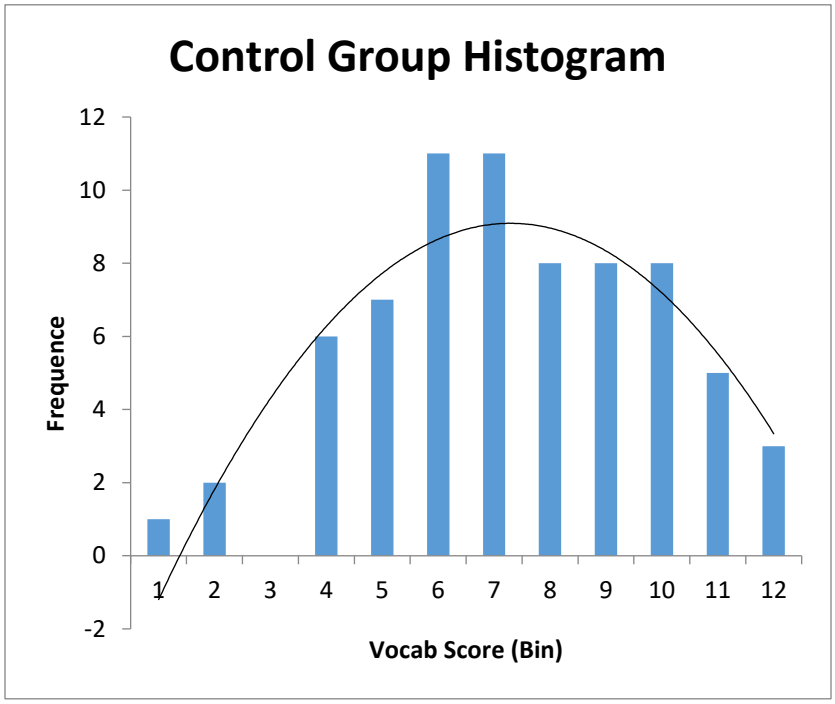
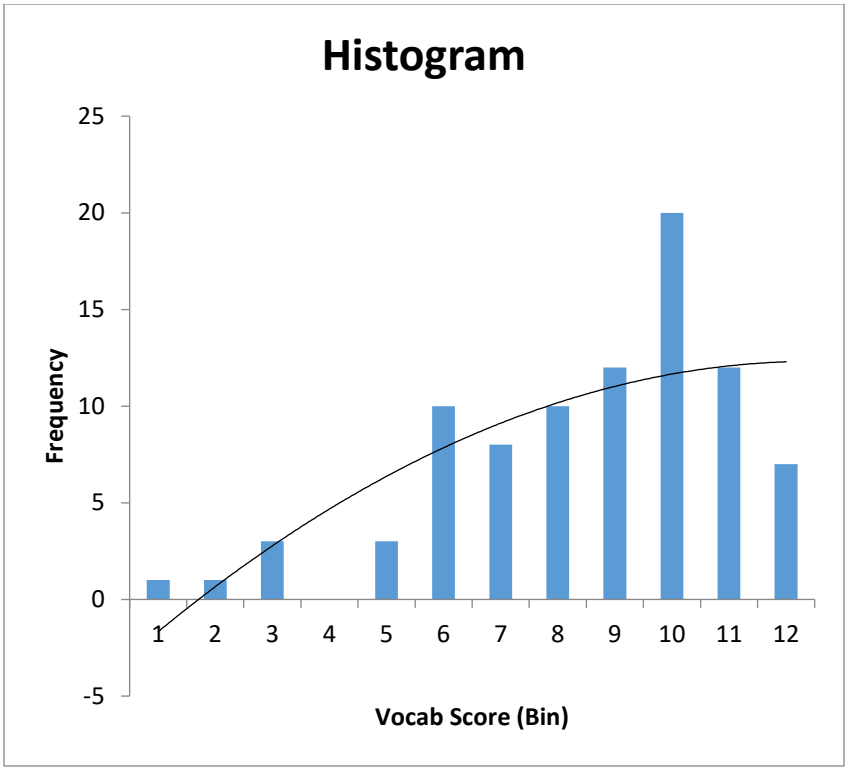


Figure 2



Coding

In the regression analysis in Table 5, data was cleaned and coded as follows. Two students had no gender listed and their data was removed. Female students were coded as 1 and male students coded as 0. P.S. Students Students were coded with 1 and control group students were coded with 0. Students speaking Non-English Languages or Other were coded as 1 and English or bilingual homes were coded as 0.

Nesting

Students in classrooms are subject to the teaching in groups, which is to say students are "nested" in classrooms. An analysis employing multilevel regression model is often stronger than the ordinary least squares regression employed in evaluation, though more complex than fitting this evaluation project. Multilevel regressions typically increase the size of standard errors. Given the size of the high level of statistical significance, a small increase in the size of the standard errors is unlikely to impact the overall statistical significance level.