**Planning and Conducting an Evaluation of an Inquiry-based Science Kit Intervention in Four Public School Districts**

**Sarah J. Brasiel, Leslie Grunden, and Eric Rolfhus, Edvance Research, Inc.**

*This paper presents the challenges and lessons learned from an external evaluation of a regional program designed to improve science and literacy achievement for high-needs students in grades K–6. The program's goals were to strengthen teacher understanding of science and quality of instruction in 4 districts, through professional development and coaching that supports inquiry-based science kit instruction. Lessons learned for evaluators include the importance of communication with project researchers to understand the day-to-day implementation issues that could potentially bias analytic findings, the challenges around access to student-level data to create propensity-score matched samples when random assignment is not possible, the use of descriptive data and graphical representations to improve clarity of the findings to local stakeholders, and documenting implementation issues throughout the study to guide researchers in improving implementation in future years.*

This evaluation is of an inquiry-science curriculum and instruction intervention for teachers in grades K–6 in 10 participating schools within 4 school districts in the Southwest. Teachers received professional development (PD) in an inquiry-based science approach during the summer 2010. The teachers implemented the first inquiry-based science kit in their classrooms during the fall 2010 semester and the second kit during the spring 2011 semester. The change between pretest and posttest levels of proficiency is analyzed to determine the extent to which teacher understanding increased following PD and whether student understanding increased following their engagement in activities and instruction with the specific inquiry-based science kit. The evaluation team learned a range of lessons over the past year, which will be discussed along with the major findings of this evaluation.

**Inquiry science project overview**

The goal of the project was to improve student learning and to close achievement gaps in science for high-need students by strengthening teacher understanding and quality of instruction in the classrooms. The following research-based instructional materials were provided to participating teachers and schools: 1) Full Option Science System (FOSS) program materials, which were designed to provide meaningful science education for all students in diverse American classrooms and to prepare them for life in the 21st century,[[1]](#footnote-1) and 2) Science and Technology for Children (STC) supplemented with the new BOOKS™ materials developed by the National Science Resource Center (NSRC).[[2]](#footnote-2) This current project provided the FOSS and STC science materials to each teacher for use in their classroom following PD.

Teachers received one week of PD in August 2010, just prior to the beginning of the school year. Teachers new to the school during the year were provided opportunities to attend a shortened PD in September 2010 and January 2011. The PD addressed how to deliver effective instruction to students using two FOSS or STC modules appropriate for the teacher's assigned grade. The project is implemented across all grades K–6; however, data is only collected for teachers and students in grades 3–6. The assessment data is collected from aligned assessments developed by either the FOSS or STC publishers, which are only available for grades 3–6. These assessments are intended for students, but were administered to teachers to assess changes in teacher understanding before and after PD and administered to students before and after exposure to the inquiry-based science kit instruction.

**Lessons learned from year 1 evaluation of an inquiry-based science kit program**

In this paper we provide a description of some important lessons we learned as evaluators related to issues and how they were solved as well as suggestions to improve the project in subsequent years. The following lessons learned will be described:

* Ensuring quality data for informative analysis
* Problem-solving data collection issues
* Presenting findings that are of utility for program improvement
* Documenting implementation issues to improve research in subsequent years
* Improving on-time delivery of work through project management

**Ensuring quality data for informative analysis**

There were two kinds of inquiry science kits: FOSS and STC. One provided only multiple choice items and the other provided some open-ended items that were scored using a rubric. Students were given 1 point, just for marking any response and zero points only if they did not respond. This scoring rubric made understanding change in student understanding from pretest to posttest challenging, because a student who marked any response would be rated higher than a student who chose not to respond at all. For the first semester of the year, as evaluators we were only provided the final total performance score for each kit for each student using this scoring system. For example, if a student took an assessment that had 20 items, and for each item they could earn up to 3 points, then the total possible points would be 60. We were provided with a data file that gave the total points each student received. If student X received 50 points, then their performance as a percent was 83 percent (50/60).

There were unusual patterns between pretest and posttest student performance in the data with students losing ground on some of the assessments, perhaps due to leaving responses blank instead of guessing. We requested item level data to better understand these issues. The project lead petitioned the board for additional funds to hire a temporary staff person to enter the item level data for analysis. The funding was approved and the item level data was made available. We re-ran all the analyses, correcting the final performance scores by subtracting one point for all items where students were awarded a point for guessing. This also provided more informative results for the item analysis in terms of the effect size of improvement by item to guide future instruction.

The student data file frequently was provided with errors, with students receiving ratings higher than the allowable rating, with teacher IDs that were not on the original teacher ID list, and other data anomalies. Time was spent to document these issues and discuss them with the project leader to come to a resolution and ensure the data was accurate for analysis. The project leader was responsive to e-mails with questions about the data, typically responding within 24 hours to either upload a new data file with accurate data or to provide information about specific students or teachers to allow the correction of the data. If care had not been taken to review the data when it arrived, the results would not have been accurate due to the significant amount of data issues.

**Problem-solving data collection issues**

The observation protocol used by the project for classroom observations is comprised of statements that are each given a rating, ranging from 1 to 5. However, if a behavior described within one of the items did not occur during the observation, the observers were directed to leave the rating blank during year 1 observations. Blank ratings result in missing data and complicated or skewed the results of the data analyses. It made it impossible to look at low, medium, and high implementers, or to determine if implementation was associated with other data gathered on teachers or students as was proposed for the evaluation. For subsequent years of the study, we recommended that a set of items be selected to document teacher and student behaviors that are expected to occur for any lesson. For these items, a rating would be required. We also recommended that subsequent years of the project include the determination of an acceptable benchmark (score) that would represent high versus low fidelity of implementation. Fidelity of implementation is important as it may provide information to explain student achievement findings and variability in student achievement across and within schools.

A state data request was submitted to the State Education Department at the start of the evaluation requesting student achievement data for all students in grade 3 (2009/10) and grade 4 (2010/11) in the state. The purpose of this data request was to create a comparison group of students similar in demographics and achievement to the students participating in the program through propensity score matching. It was anticipated that this requested data could be used in the evaluation of the effect of the inquiry-science program on student achievement. This data request was denied in March 2011 due to changes in state administration and the data management system. Therefore, the project leaders obtained data from schools in participating districts who were not implementing the FOSS and STC kits and one additional district to be used for the propensity score matching. This was the best alternative, given the lack of availability of state data.

**Presenting findings that are of utility for program improvement**

Presenting findings in tables with significance levels was not informative for program improvement. Neither were typical graphs to describe pretest and posttest differences (as shown in the figure that follows).

Figure 1. Distribution of student performance (percent) at pretest and posttest, spring 2011

 

The client provided feedback that the main users of these reports were not as interested tests of statistical significance (p-value, effect size), or all of the graphical displays. Although the users were very pleased by the progress the students had made, they wanted something more to guide support for teachers and inform improvements to the program the following year. We brainstormed ways to provide item level information that could be used by the science coaches to understand strengths and weaknesses in teacher and student understanding to direct their support.

What we ended up with was an *Item response analysis*, where the results of the item analysis were provided directly on the assessment. Items with a small effect on the item performance analysis were identified with a symbol. These items represent areas of need for additional instruction to improve student understanding. What follows is an example of this type of information provided to inform the work of the science coaches

This symbol is used for items with small effect sizes for students.

**T**

**S**

This symbol is used for items with small effect sizes for teachers.

## 

## Item Response Analysis Example

se68. What causes the 24-hour cycle of night and day?

The rotation of the Earth.

**T**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Response** | **Percentage** | | | |
| **se68.** | **If the student...** |  | *(Rating)* | **Student**  **Pretest** | **Student**  **Posttest** | **Teacher**  **Pretest** | **Teacher**  **Posttest** |
| 2 | indicates the rotation of the Earth causes the 24-hour cycle of night and day. |  | **2** | 38 | 68 | 90 | 75 |
| 1 | provides any other answer. |  | **1** | 48 | 27 | 10 | 25 |
| 0 | makes no attempt. |  | **0** | 15 | 5 | - | - |

We also provided reflection worksheets to be used by the Science coaches with the item level data. An example of a reflection worksheet for the Sun, Moon, and Stars kit is shown below. The client responded positively to this item analysis report with reflections and used them in the science coach training the summer prior to the second year of the program. The project lead reported that these reports were well received, and the science coaches were able to use them to target areas in need of additional support for teachers and re-teaching.

# Reflections

For each inquiry-based science kit, there is a final section to allow time for reflection after reviewing the student and teacher data for the kit. A few question prompts are provided to direct the reflection. These can be used in small group discussions with teachers as part of the science coach\_ support during the year.

1. Take a look at item Se 68. This item had a small effect for teachers. What makes this item difficult? What instructional strategies might be used to improve understanding related to this item?

For the main analysis using HLM to account for clustering of students in classrooms and student and teacher factors that might explain some of the differences in performance in the state science assessment, we provided a graphical display of the concept of how different factors at the student level and at the teacher level explain some of the variance in the student state science assessment scores. The more technical table we provided in an appendix. The figure that follows is an example of this more user friendly figure provided to better understand the findings. This figure was accompanied by text that described the teacher and student factors that were statistically significant in the model.

**Documenting implementation issues to improve research in subsequent years**

Weekly or biweekly conference calls were held with the project leader to better understand implementation and data collection issues. It was important to provide a detailed agenda with an opportunity for the project leader to provide updates on whether the project was on schedule and to provide us with data; in addition, we provided updates on where we were at with our schedule of evaluation deliverables. This communication ensured advance notice of any delays and allowed for planning of longer periods of data analysis and report writing as needed.

We also set up, as part of the statement of work, a plan for interim reports prior to the end of year final evaluation. A teacher assessment performance report summarized the performance of teachers on the kit assessment before and after PD. Next, a student and teacher demographics report was provided based on the data collected from schools. Fall and spring student performance report were provided with detailed overviews of the performance by kit, by district, and by classroom. These reports informed the work of the Science Coaches with the teachers and informed the project leaders and the funder’s board of the progress of teachers and students in the project.

Another reason for the importance of communication was periodic updates related to implementation that could bias the findings. For example, an implementation issue was when the project leader explained that one of the observations was conducted after the teacher had completed the kit, so the observation was of the traditional curriculum. In response to this update, we removed this data from the analysis, since the focus was to be on in-class implementation of the inquiry-science kit curriculum and instruction. In essence, the evaluation benefited from this kind of open communication with the client.

In addition, we provided recommendations for future years of the project in terms of improving data collection, adding rigor to the evaluation design, and improving data accuracy and data management. We held conference calls with the client about these recommendations asking the client to make decisions about changes to the year 2 program that were feasible and those that would need further consideration. Based on this information, we designed a plan for evaluation for the second year of the program.

These communication efforts helped the client feel that we were on top of issues that needed to be addressed. It also gave them confidence in our ability as researchers to improve on the study design. The interim reports allowed them to see the progress and get additional funding to begin scale-up of the program.

**Improving on-time delivery of work through project management**

The original statement of work included hard dates for deliverables based on exact dates for receipt of data. Unfortunately, data was never received by the date in the statement of work, putting pressure on us to provide the analysis and report of findings quicker than anticipated. The project manager worked with the client to amend the statement of work to allow for more flexible deliverable deadlines within 20 business days following receipt of data to allow for sufficient time for analysis and reporting when data did not arrive on time.

1. See <http://lawrencehallofscience.org/foss/introduction/index.html> [↑](#footnote-ref-1)
2. NSRC launched Science and Technology for Children BOOKS™ to add a much-requested literacy component to its popular curriculum for elementary students. Each book is designed to be used in conjunction with teaching the STC unit of the same name or as a stand-alone resource that conveys topics in a way that makes science interesting and relevant. See <http://www.nsrconline.org/curriculum_resources/science_readers.html> [↑](#footnote-ref-2)