LONGITUDINAL EXPERIMENTS IMPROVE UNDERSTANDING OF NATURE OF SCIENCE

Longitudinal Experiments in a Middle School Life Science Class Improve Understanding of the Nature of Science

Rhonda Hoenigman
University of Colorado

Sandra O’Donovan
Louisville (CO) Middle School

Sarah Hug
University of Colorado

Author Note
Correspondence concerning this article should be addressed to Rhonda Hoenigman, University of Colorado, Department of Computer Science, 430 UCB, Boulder, CO, 80309. Email: rhonda.hoenigman@colorado.edu
Abstract

A middle school life science teacher performed an action research study to determine if longitudinal experiments in her seventh-grade life science class improved her students' understanding of experimental design and analysis on constructed response items. Students engaged in three inquiry-based, longitudinal labs over the course of the year as background tasks to the regular curriculum. They designed experiments and collected and analyzed experimental data individually and as part of classroom discussions. At the end of the school year, a test was administered to her classes and a control group that measured students' abilities to analyze an experimental design and develop a follow-up experiment. Students outperformed the control group on this exam by a statistically significant margin. Scores from this exam showed that these longitudinal labs engaged both the higher- and lower-performing students. This paper presents this approach for incorporating longitudinal science experiments into a life science classroom.

*Keywords:* nature of science, inquiry-based science, longitudinal life science experiments
Introduction

The Nature of Science (NOS) is difficult to define in a few words. It has been stated that including the NOS in science education means that the curriculum needs to include more than a collection of isolated facts. It needs to include a discussion “about the character of scientific knowledge, how it has been developed, and how it is used.” (Hurd, 1960). This approach to the NOS matches the activities of real scientists, where considerable effort goes into experimental design and analysis, and experiments are designed to reveal new information about the world. In this context, there is not one best way to do science, experiments can take many forms with differing objectives (AAAS, 1990; Dunbar, 2000).

Unfortunately, experimental design and analysis are often not the dominant instruction in most K-12 science (Fortus, 2005). Lab activities are more likely to include a pre-defined set of steps, with known outcomes and designed to fit into a 50-minute class period; it is not uncommon for students to describe what they know was supposed to happen instead of what they observed. The emphasis is on getting the “right” answer instead of understanding the concepts being taught, and experiments that don't support the hypothesis are viewed as incorrect (Lord, 2006; Thompson, 2010). Inquiry-based learning activities are designed to remove some of this formulaic structure by encouraging students to pose research questions and then uncover answers to these questions (NRC, 1996). This approach is seen as a way to educate students on the NOS as they perform the same activities as scientists. While inquiry-based activities have been shown to improve students understanding of the NOS (Marx, 2004; Zohar, 2002), there can be several challenges to implementing this approach as part of a curriculum that reaches the average student in the average classroom. Classroom activities may require expensive lab materials, software, field trips, or teacher training (Bouillion, 2001; Stratford, 1998).

There is considerable research showing the effectiveness of inquiry-based activities as part of summer science camps for increasing both interest in and knowledge of science, such as the study presented in Gibson (2002). Students who attend these camps are often self-selecting, i.e. they are interested in science and therefore choose to go to a science camp. Students who are disinterested or only moderately interested in science or who might be facing other family constraints do not attend a summer science camp. However, students in these groups can represent be a significant portion of a school district's population, and a large determinant of scientific literacy and standardized test scores. Reaching this population and increasing their knowledge of the NOS could have a significant effect on scientific literacy.

One type of science experiment that is difficult to include in even the best inquiry-based lessons is a longitudinal study – an experiment that occurs over time. Professional science experiments often occur over weeks, months, or years, and scientists collect data for the duration of the experiment. In a classroom setting, where the curriculum dictates that every minute is filled, it can be difficult to allot time for a lengthy experiment on one topic. In some classrooms, longitudinal experiments are incorporated into the curriculum through a science fair project. Students work independently or in small groups to design their own experiment outside the classroom setting. However, this approach also presents problems. While some students can excel without one-on-one mentoring from the teacher, many students are lost in the process. It is difficult for a teacher to have 100 meaningful conversations with 100 students all doing different science fair projects. In the experience of the teachers involved in this effort, many students generate bad, untestable hypotheses under these less structured conditions and conduct poorly designed experiments. When these experiments do not work, they often do not know why. Even
when the experiment is successful, some students can have a difficult time explaining what happened and why.

In this paper, we present an inquiry-based approach that addresses the issues often encountered in inquiry-based activities and incorporates the element of time through a series of longitudinal lab activities. The lab activities are extensions of the regular curriculum that can be run as background tasks, thus allowing regular instruction to continue on schedule while still incorporating the NOS into the curriculum. This approach is not dissimilar to real scientific longitudinal studies in which scientists collect data at intervals to get a snapshot of how the system is changing, and between these data-collection events, other tasks take priority. Students perform three longitudinal lab experiments (approximately six weeks each) where the focus is on experimental design and analysis in three different topic areas. These topics are covered in lectures alongside the experiments. Incorporating the NOS into the curriculum using multiple topic areas (in this case, single-celled organisms, plants, and worm bins) allows students to see the broader nature of experimental design and analysis presented from different angles. Focusing on different topic areas also draws in more students, i.e. some students are more interested in plants than single-celled organisms, and vice versa. With each experiment, we also increase the students' responsibility in the experimental design and evaluation of the experiment and the results. By covering these concepts as part of class discussions about each lab activity, we are able to provide scaffolding for students who need additional help, and get consistent engagement from students of all levels.

In our implementation of this approach, we found that students were engaged in the labs throughout the year and showed a noticeable improvement in their understanding of experimental design and analysis. The student engagement and buy-in that we observed in these labs supports previous research on the value of project-based science education (Blumenfeld, 1991). By the end of the year, students were able to brainstorm several testable hypotheses when presented with a problem. They were also able to identify sources of experimental error when experiments did not produce expected results.

Method

The Research Context

The teachers. “Sandy” has been teaching for fifteen years, including eight years teaching seventh and eighth grade science at a middle school in the western United States. She has a Bachelor’s degree in Environmental, Population, and Organismic Biology and a Master’s degree in Education. Prior to teaching, she worked in the healthcare and biotechnology industries.

“Rhonda” is a PhD student studying modeling and optimization in a computer science department at a large research university. Rhonda was assigned to work in Sandy’s classroom as part of the GK-12 program that places graduate students in K-12 classrooms to introduce students to scientists (Thompson, 2010). The objective of this particular GK-12 program is to bring more computer science and computational thinking into traditional education. For the purposes of this project, computational thinking is defined as approaching a problem in a procedural, mathematical way in which solutions are constructed from a series of operations on quantitative data (Hambrusch, 2009; Lu, 2009).

When Sandy and Rhonda first started working together, Sandy wanted to increase the amount of experimental design, data collection, and data analysis in her curriculum. Both Sandy and Rhonda felt that incorporating longitudinal experiments into the curriculum was the best way to accomplish this goal. Having a computer scientist in the room provided the impetus to design
labs that included these concepts. Rhonda saw the focus on these concepts as building blocks to computational thinking that were necessary for understanding more complicated computational problems.

**The school.** The research was conducted at Louisville Middle School (LMS). The school population is primarily Caucasian students (80%) and Hispanic students (12%) with a relatively small proportion of students on free and reduced lunch (13%). Scores at LMS on the State Student Assessment Program (CSAP) are above average for the school district and the state. Improving CSAP scores in the areas of experimental design and analysis has been an area of focus for the school. The science standardized test is given in eighth grade and includes equal content from all three grades. About 25% of the test covers experimental design and analysis.

**The study context.** In 2009, the school district adopted the National Science Resource Center's curriculum – Science and Technology Concepts for Middle Schools (STC) – in addition to traditional textbooks as an option for middle school science. The curriculum is designed to be hands-on and encourage inquiry-based learning. LMS choose to use a hybrid of the STC curriculum and textbooks.

During her first year working with the curriculum in her classroom, Sandy implemented several of the labs as prescribed. While she was basically happy with the labs, she also felt that they often did not go far enough to encourage formal data collection (quantitative or qualitative) and analysis and a formal lab write-up. The labs included a few concluding questions, but did little to encourage students to communicate their experimental findings in writing to synthesize the complete lab. There was also very little focus on how experimental design could affect the results.

We modified a few of the STC labs to address these concerns, allowing us to keep the inquiry-based setup already present in the lab while incorporating the formalism that Sandy felt was missing. There were two labs in particular – a lab in which students set up a miniature pond and observe it over time, and a plant lab where students grow plants to observe their phenotypes – that we felt could be modified to include more elements of experimental design and analysis. We also included graphing as part of data analysis. Graphing skill is part of the math standards, and incorporating this in science class provided students with the opportunity to practice a skill that they had learned in their math classes. This skill also encouraged computational thinking – students learned about quantitative representation of data.

**Pond lab.** The first lab we modified was the pond lab, where students use rocks, dirt, plant life, and water from a nearby pond to build a small model pond. This lab paralleled classroom instruction on the definition of a living thing, microscopes, single-celled organisms, and cellular functions. As presented by STC, the lab included student observations only at the beginning and three weeks later at the end of the lab. We modified the lab to include twice-weekly observations and data collection for six weeks. We used the STC curriculum lab setup, which required one class period.

Our approach to this lab diverged from standard procedure where students begin every lab with a hypothesis. Students observed their ponds a few times over the course of approximately one week, and then Sandy led a class discussion about potential hypotheses for the lab based on students' initial observations. She asked questions such as, “Do you think you will see more organisms over time?” and “Do you think the types of organisms will change?” We used this approach for a few reasons. This process helped students understand the lab so that they could generate a good hypothesis. We wanted students to think critically about what they were observing in the ponds and avoid a hypothesis that offered limited scientific merit, such as, “We'll see living things.”
LONGITUDINAL EXPERIMENTS IMPROVE UNDERSTANDING OF NATURE OF SCIENCE

Following the class discussion, most students used the hypothesis that the class brainstormed, and a few students used their own hypotheses. This process of questioning and hypothesis generation modeled how to generate a testable hypothesis – providing scaffolding for future projects where Sandy was less hands-on during this process.

To conduct the lab, students worked in small groups to collect samples from their ponds and observe them under the microscope for approximately 15 minutes twice a week. They generated rough estimates of how many of each type of critter they observed. After two weeks, we evaluated students' data to make sure everyone was on track. We checked that they were recording their observations each week, and that their observations were what we expected, i.e. they were collecting data on the numbers and types of organisms. After six weeks of data collection, Sandy showed the students how to graph their data. For most students, their graph showed time on the x-axis and number of each species type on the y-axis. These graphs were included in a lab write-up where students discussed their results. Sandy stressed to students that the graph they generated needed to be relevant to their hypothesis. Students also discussed how to improve the lab in response to some of their results.

**Plant lab.** The second lab that we modified from the STC curriculum was about plant growth and Mendelian genetics. In this lab, students grew Wisconsin Fast Plants©, aka *Brassica rapa* from seed to observe dominant and recessive traits in the population and the passing of traits from one generation to the next. The lab paralleled instruction on Mendel's pea experiments and life cycles. We wanted this lab to repeat the concepts of experimental design, data collection, experimental error analysis, and data analysis that we covered in the pond lab using plant growth as our subject area. We modified this lab to have students measure plant growth and water use as the plants were growing.

Sandy began the lab by brainstorming with students possible hypotheses for their experiments. Since most of the pond lab data had been qualitative only, she also differentiated and explained what types of quantitative data could be collected. With the pond lab, there were also many unknowns in how the ponds would perform and the types of things we could observe, so it made sense to observe the ponds before beginning formal hypothesis testing. With the plant lab, the question was much more straightforward – we wanted to know how water use would change with plant growth. Class discussions familiarized students with the plant lifecycle. Some students hypothesized that water use would increase as plant size increased. Others believed that water use would be initially high while the plant was establishing, and then decrease later. Students also had many suggestions for how to measure plant growth, including plant height, number of leaves, number of flowers, and size of the leaves. In this lab, more so than in the first lab, students designed their own hypotheses and chose which data about plant size they were going to collect to test them. Students again worked in small groups and set up the plant experiments using the equipment provided in the STC curriculum. Plants were in individual cups, each with its own water supply, which made it easy to measure how much water each individual plant used between measurements. They measured water use twice a week for six weeks by recording the drop in the water level from a known amount. Following data collection, students graphed their data (growth vs. time and water use vs. time) and included this information in their lab report.

**Worm bins.** For the third longitudinal experiment, we did not use the STC curriculum, but designed a lab ourselves. We wanted a lab to parallel classroom instruction on communities, ecosystems, kingdoms of living things, and the role of decomposers in ecosystems. We designed an activity studying composting worm bins that covered the concepts of experimental design, data collection, and data analysis, as well as the relevant classroom subject matter. Worm bins
are constructed from dirt, newspapers, and special composting worms, called red wigglers (Grossman, 1997), and are used for composting household food scraps. Food is added to the bin and consumed by bacteria, which is food for the worms. Over time, the worm population increases, producing compost (worm casings) that can be removed from the bin and used in the garden. In addition to the curriculum mentioned above, this experiment gave us a chance to talk about bacteria, biotic and abiotic elements (revisited from the pond lab) and energy flow through ecosystems.

The environmental conditions and the food types in the bin affect how quickly the worm population grows. A bin that is too hot or cold, too wet or dry, or does not have the right types of food may not see as much worm growth as a bin with ideal conditions. These variables provided the basis for our experiment. Sandy introduced worm bins and how they work, and then, following the pattern established with the first two labs, students brainstormed several ideas for experiments with the worm bins. Some of these ideas included setting up bins with different food types, temperatures, or moisture levels. After all ideas were presented, students voted as a class on two ideas to pursue for experiments. Each class had three bins – one control bin and one bin for each experimental condition. We had four existing bins that we used to seed the new bins by dividing each bin into three bins, and then adding additional dirt and newspapers to each bin. The control bin was placed in good conditions (sunny), and fed a variety of fruit and vegetable food scraps. In each experimental bin, one variable, such as the type of food or environmental conditions, was modified. For example, in one class, students modified sun exposure for one bin and moisture level for the other bin. They changed the sun exposure by placing the bin in the shade. They modified the moisture level by not adding water to the bin when they added it to the other two bins. We used leftovers from the school cafeteria salad bar to feed the worm bins. Twice a week, randomly selected students added food to the bins, and collected data on the amount of food remaining that had previously been added, the temperature in the bin, and the worm count. This data collection took about 15 minutes, just as it did with the pond and plant labs.

The worm bins presented a difficulty that we did not experience in the first two labs. We needed to determine how the population of the worm bin was changing. However, removing all the worms and counting them each week was not an option due to the time involved, stress on the worms, and the overall disgust with sorting through worms and rotten food. Many students had a difficult enough time stirring the worm bins with gardening tools, let alone touching the worms with their hands. We considered having students stir up the bins, and then scoop out a cup of dirt and count the worms in the sample. However, even this process would take more time than our scheduled 15 minutes allowed. We decided instead on a method of counting clumps of worms. It is common in worm bins for the worms clump together, and in any worm bin, there can be multiple clumps. To determine worm count, students stirred the bins and counted the clumps. They also counted clumps of baby worms as a leading indicator of how the population was changing.

**Existing standardized testing.** In March of each school year, every seventh grader at LMS takes a spring science assessment, which is used as a placement exam for entrance into advanced eighth grade science. The exam consists of a publicly available CSAP science question and a few math questions. We wanted to evaluate if our longitudinal labs had a measurable effect on students’ scores on this exam when compared to students from other seventh grade science classes at LMS. The CSAP question included four parts about the following fictional scenario:

1. Ralph's teacher gave him a slice of bread with small spots of mold on it to do an experiment on mold. To do the experiment, Ralph used tracing paper that had a square
grid printed on it. He placed the paper on the slice of bread and traced the bread outline. There were 100 squares covering the whole slice of bread, and 2 squares covered mold spots. Ralph counted the number of squares covering the mold spots every day for 5 days. The students were then asked to write one question that Ralph was trying to answer.

2. Students were given a data table and asked to draw a graph of the data, then use the graph to estimate mold growth on the next day.

3. Students were provided with this description: Ralph decided to do another experiment. He took three slices of bread, each having mold spots covered by 2 squares. He put one slice in the oven at 200°F, one slice on the kitchen counter, and one slice in the refrigerator. The students were then asked to write one hypothesis for Ralph's experiment.

4. Finally, students were provided with fictional experimental data and asked to write one conclusion that Ralph could draw from his data.

**Research Design Assessment**

In addition to the spring science assessment, we wrote our own research design assessment that specifically addressed the concepts we had covered in the longitudinal labs. Our assessment served two purposes: to assess the concepts we were teaching in our labs, and to serve as a possible revision of the spring science assessment. In past years, several students who performed well on the spring assessment performed poorly in advanced science, possibly suggesting that they had just learned how to take standardized tests, but did not really understand the material. Both Sandy and the control group teacher were interested in revising the exam to include questions to assess how well students could think about abstract NOS concepts. We used the same scenario as presented in the spring science assessment so that all students were familiar with the question. However, we asked questions related to experimental error, effects of error, and experimental design and analysis. We presented students with the following scenario:

Ralph's teacher gave him a slice of bread with small spots of mold on it to do an experiment on mold. To do the experiment, Ralph used tracing paper that had a square grid printed on it. He placed the paper on the slice of bread and traced the bread outline. There were 100 squares covering the whole slice of bread, and 2 squares covered mold spots. Ralph wanted to determine the best growing conditions for mold. He took three slices of bread, each having mold spots covered by 2 squares. He put one slice in the oven at 200°F.

Students were presented with the following five questions:

1. Identify one source of error in Ralph's experiment.
2. How could this error affect Ralph's results?
3. Design a follow-up experiment to address this error.
4. At the end of three days, there was no difference in mold growth between the three environments. What should Ralph do?
5. At the end of the experiment, Ralph did not have data to support his hypothesis. What should he write in his lab report?
6. Even though the experiment didn't go as planned, what could Ralph learn from his results?
Assessment administration. The exam was administered to all students from both groups on the same day during the last week of school. We attempted to administer the test consistently to both the test group and the control group. Students were told that the second assessment was being given to help make a more accurate science assessment. In Sandy's classes, students were also told that the test was to evaluate the longitudinal labs. Approximately 200 seventh graders took the test, split evenly between the two groups.

Assessment scoring. To score the research design assessment, one point was awarded for each question. We treated the first three questions as a group. We awarded points for Questions 2 and 3 even if Question 1 was incorrect, provided that the answers for Questions 2 and 3 followed logically from the reasoning in Question 1. Common answers to these questions from both Sandy's students and the control group were about the time length of the experiment, other variables, such as humidity, inconsistent conditions that could have affected the results, such as humidity, inconsistent conditions, and the inability to get an accurate count of the mold. As an example, the following answers that referred to light being an additional variable in the experiment received a perfect score.

- Identify one source of error in Ralph's experiment. Answer: The slice on the counter would probably get more light than the other two slices.
- How could this error affect Ralph's results? Answer: There would be two variables instead of one for the slice on the counter.
- Design a follow-up experiment to address this error. Answer: He could do the same thing, except for one difference. There would be no light getting to the mold on the slice on the counter.

Another student also answered that other variables could affect the results. However, his answer did not include a good follow-up experiment. He received a point for the first two questions, but not for the third question:

- Identify one source of error in Ralph's experiment. Answer: His experiment wasn't controlled, he didn't control air moisture or light.
- How could this error affect Ralph's results? Answer: Ralph's results wouldn't be solely on the temperature.
- Design a follow-up experiment to address this error. Answer: He could put the cold bread in a bucket of ice, and the hot bread by a space heater.

Other students cited the inconsistent conditions on the counter as a source of error. The following answers received full credit:

- Identify one source of error in Ralph's experiment. Answer: The slice of bread on the counter would not always be at the same temperature. Ralph might decide to turn on the heater some days or open up the screen door.
- How could this error affect Ralph's results? Answer: The rate of growing mold would change either faster or slower.
- Design a follow-up experiment to address this error. Answer: Ralph could put the slide of bread that he put on the counter in a temperature-controlled box or room.
- Note: Presumably, room temperature should never be as warm as 200°F or as cold as the inside of a refrigerator. However, we decided to ignore this nuance and accept this line of reasoning. Another student commented that, in the varying temperatures, the mold may grow better or worse than it would at a steady temperature. This answer addresses the changing temperatures as the variable, rather than the actual temperature range.

Other students mentioned that it might be difficult to get an accurate measurement, due to the grid and other factors. One student answered:
Identify one source of error in Ralph's experiment. Answer: An error in this experiment is that there was no exact measurement for mold growth.

How could this error affect Ralph's results? Answer: It may have seemed that there was no difference, however, you can't do a spore count/exact area count by looking at break through a piece of vellum paper. Also, mold in some spots may be on the surface, while all the way through in others.

Design a follow-up experiment to address this error. Answer: He should wait five days and only take data at the end. Once it is done, he can cut it open and more thoroughly measure the area of the mold.

Another student focused on the grid to support his answer that there could be measurement errors.

Identify one source of error in Ralph's experiment. Answer: Mold doesn't grow in exact cubic centimeters so it's hard to measure.

How could this error affect Ralph's results? Answer: One slice of bread might start with more so it is likely to have more at the end of the experiment.

Design a follow-up experiment to address this error. Answer: He could start with identical pieces of bread and then take some amounts of mold from an already moldy piece of bread and distribute them evenly on the three pieces of bread.

Scoring difficulties. The open-ended nature of these questions also made them difficult to grade. In many cases, it was difficult to assess students' intent. We did not know if they had good reasoning, but difficulty articulating their answer, or if they were reaching for a generic response just to write something on the page. At this stage in their academic careers, students know what a correct answer is supposed to look like. They could write an answer that sounded good, but they were really just grasping at straws. For example, one student answered the first three questions in the following way:

Identify one source of error in Ralph's experiment. Answer: Putting a piece of bread in the oven for five days at 200°F.

How could this error affect Ralph's results? Answer: He burns the bread.

Design a follow-up experiment to address this error. Answer: Lower the temperature of the oven.

In this case, the student's answers were not well articulated. We awarded the student points for the second and third answers, but not the first because putting a piece of bread in an oven for five days is not necessarily an error. However, his conclusion that this could burn the bread is a legitimate experimental error. His answers were not well articulated overall, which also made his exam difficult to grade.

In contrast, other students were more articulate, but it was difficult to determine if there was any substance in their responses. For example, one student submitted these responses:

Identify one source of error in Ralph's experiment. Answer: One error is that the mold may not be in the same spot on all three slices.

How could this error affect Ralph's results? Answer: Some places could grow more mold than other parts on the bread.

Design a follow-up experiment to address this error. Answer: He could find bread with mold in the exact spot as other slices.

If the answers to questions two and three had been more specific, we would have also awarded a point for question one. Since we did not see anything specific, it was difficult to determine if the student had a clear idea he was trying to convey, or if he was providing a generic answer. As a
result, we did not award any points for his answers. It was unclear from reading his answers why the part of the bread where the mold grew was significant.

Several students responded that the position of the mold could affect how well it grew. Students who received points for these answers were able to articulate ideas for how the position could affect growth, such as mold spots next to each other might grow faster than mold on different parts of the bread.

For the other three questions, we encountered similar difficulty with deciphering the level of understanding in students' answers. However, there was one notable difference between Sandy's students and the control students – Sandy's students attempted an answer for all questions, and not one student answered a question with “I don't know” or “I don't care”. This result perhaps speaks to the level of personal attachment that her students felt to these labs.

Results

We evaluated how students performed on the spring science assessment and our research design assessment to determine if the longitudinal labs had a measureable effect on students' scores on these exams. We compared the mean and median scores from Sandy's students to scores from other seventh grade science students (our control group) for both assessments. On the spring science assessment (the CSAP question), our scoring for each question was based on the rubric provided with the assessment. The maximum possible score on the test was 14. The means for Sandy's students and the control students were 11.23 and 11.44 respectively, with standard deviations of 2.01 and 2.36. Using a two-tailed, two-sample test, we found no statistically significant difference. The medians for the two groups were 11 and 12, indicating that the scores from the control group were slightly higher. Histograms of the score distributions for the experimental and the control group are shown in Figures 1 and 2. These images show that the scores for both groups appear to have a non-normal distribution, which makes the mean and standard deviation an imperfect comparison for the populations.

![Figure 1](image.png)

*Figure 1. Sandy’s students’ scores on spring science assessment.*
LONGITUDINAL EXPERIMENTS IMPROVE UNDERSTANDING OF NATURE OF SCIENCE

On our research design assessment, the maximum possible score on the exam was 6. The mean for Sandy's classes was 3.97, with a standard deviation of 1.43, and the mean for the control students was 3.52, with a standard deviation of 1.45. We applied a two-tailed, two-sample test and got a p-value of 0.035, a value that is within the 95 percent confidence interval, and sufficient to support claim that there was a difference between the two groups. Based on the results from this exam, there was a statistically significant difference in students' ability to answer questions on the topics we covered – experimental design and analysis – between the classes that included our longitudinal labs and those that did not.

Both populations had a median score of 4. The score distributions for this test for the experimental and the control group are shown in Figures 3 and 4 respectively. For the control group, the population is normally distributed with a clear peak at the mean and a slight skew to the left (below) the mean. For Sandy's classes, the scores are clearly not normally distributed. Rather, the entire population is shifted to the right, meaning that both the lowest and the highest scores were improved over scores in the control group. The lowest scores appear as outliers in this distribution, given the clustering of the other scores. We recognize that our method of administering the exam may have introduced some bias. We told students that the exam was to evaluate the success of the lab and Sandy's students may have been more motivated to take the exam seriously due to their connection to the labs and us. However, anecdotally, we believe that the longitudinal experiments engaged students who normally do not perform well in science and kept them interested in science throughout the school year. The score distribution supports this observation – there are fewer low scores than in the control group.
The distributions on these assessments bring up an issue that both teachers had anecdotally observed. All of Sandy's classes were held in the morning, while the control classes were all in the afternoon. Due to scheduling of other advanced and special education classes, morning science classes typically had a more uniform student population than afternoon classes. In the CSAP assessment, this could explain the score distribution for Sandy's students, where there are fewer perfect scores and few scores in the lower range. We see the same uniformity in our assessment – scores are consistently shifted to the right. It is also worth noting, however, that more students earned perfect scores on this exam than in the control group, which is opposite of the pattern observed in the CSAP assessment. We believe that this approach not only drew in low-performing students, but also increased scores for the high-performing students.

In addition to the overall performance, there were interesting results from individual students. There were several students in Sandy's class who were not typically high-performing students and did not apply to advanced science. There were a variety of reasons for their...
LONGITUDINAL EXPERIMENTS IMPROVE UNDERSTANDING OF NATURE OF SCIENCE

difficulty, including learning disabilities, poor behavior, and/or general lack of interest in science. However, a few of these students who did not do well in school earned perfect scores on our assessment, and demonstrated an unexpected level of abstract thinking. In contrast, there were students with very good grades who struggled to articulate the higher-level thinking needed to answer the questions. We hope that this type of assessment will enable us to identify actual student understanding and insight rather than just measuring how well a student listens and does what they are told. Research to identify whether this particular test was a good predictor for success in advanced science is on-going.

Lessons Learned

There were several lessons learned during these labs about teaching experimental design and the teacher mindset needed to make this approach work in the classroom. In these labs, we focused on graphing as the primary method for data analysis. After the pond lab, which was the first lab we did, Sandy drew a line graph on the board as a demonstration for how students would generate their own graphs from their data. To our dismay, about 10% of the students did not know what to do with their own data, so they just copied her sample graph into their own lab reports. They did not recognize the connection between the data they collected and the graph – that the graph was a visual representation of their data. After the plant lab, Sandy repeated this graphing review and emphasized the meaning of the graph. We saw an improvement in students' understanding, and this time, only one or two students, out of 100, did not understand the concept.

The most difficult part of these labs was monitoring that students were actually doing observations to the necessary level of detail – numbers, pictures, written descriptions. On the pond lab, we had a group of students who were feverishly looking through microscopes and proudly announcing the critters they were seeing. However, they were not writing anything down, and we did not realize this until half way through the data collection period. It is tedious to check everyday that students are staying on task. We instituted a mid-point check to identify that students were collecting sufficient data for their lab reports.

From the first to the third lab activity, we saw an increase in students' ability to design experiments and think about experimental error. We believe this was due to repetition. It was not enough to introduce a lab and expect students to generate their own testable hypothesis the first time. This process needed to be modeled by the teacher through classroom discussions brainstorming testable hypotheses, and repeated. On the first lab, most students used the hypothesis that Sandy provided. In the plant lab, came up with their own hypotheses and often collected data for multiple things, such as water use vs. plant size, and water use vs. number of leaves. By the worm bin lab, the final activity, students were able to brainstorm their own hypotheses and present several options for experiments that we had not considered ourselves. The open-ended nature of these activities meant that higher-performing students were not stifled by this repetition – they were able to generate more advanced hypotheses and data collection practices.

The worm bin exercise became, unintentionally, an exercise in experimental error and the importance of a good experimental design. We started the lab in the spring. Unfortunately, we had a cold spring and there was minimal worm growth in any of the bins. We also introduced potential error by having different students collect data each time, which resulted in inconsistent data collection methods and data. Even with these issues, the worm bin activity was a huge success from the perspective of analyzing experimental design and error. At the end of the experiment, when we looked back at our data for each week and noticed that not much had
changed since the beginning of the experiment, and that the results were inconsistent from week to week, we asked students why they thought the experiment did not work. They were able to come up with several good reasons on their own. They could identify that there were variables out of their control, such as the weather, that affected their results. They recognized that different people were estimating the worm populations differently. They were also able to come up with several improvements to the experiment to generate more consistent data. We don't believe they could have done this at the beginning of the year, or even after the first longitudinal lab. It was also a new experience for them to say that the results were inconclusive. This “failed” experiment also provided them with good “real-life” lesson – an incorrect hypothesis does not mean you didn't (or can't) learn something.

It was also interesting for us, as instructors, to hear what students thought about the labs. On the unit test, Sandy asked students to compare and contrast the pond and plant labs. She expected students to comment on experimental design and qualitative and quantitative analysis. One student said, unexpectedly, that in both labs, they were taking care of living things. This answer showed a personal attachment to the process. We also saw this attachment when students showed up for class each day. Unprompted, they would go straight to their experiments to check on their plants or their ponds.

One of the most important lessons learned is about the mindset of the teacher conducting these labs. In the pond lab and the worm bin lab, in particular, we as the instructors did not know how well the lab was going to work. We knew enough about both systems to predict that there should be changes in the populations. However, both systems are dynamic, making it difficult to predict exactly what would happen under the conditions provided. As a result, we were experiencing the labs for the first time right along with the students and were just as excited as they were about the results. In addition to excitement, this unknowing about the lab outcome can also come with some discomfort for the instructor. In labs where outcomes are already known, the teacher can be the expert in the nuances of the activity and the potential pitfalls that students might encounter. In the labs we have designed, the teacher guides the students through the process of experimental design and analysis. On these things, the teacher can be knowledgeable. However, it is not possible for the teacher to know everything about every student experiment. This brings up a point that scientists know all too well – science is often about feeling uncomfortable (Schwartz, 2008). In this way, our longitudinal labs incorporate another element of real science experiments – the element of not always knowing the right answer and the ability to be comfortable in this state.

**Year Two:** The information presented in this paper covers the first year of implementing these labs. At the time of this writing, the pond and plant labs have been implemented a second time. Some of the issues we encountered the first year were encountered again, and other elements of the labs ran more smoothly the second time around. In the pond lab, we experienced similar problems with inconsistent data collection. Students were engaged in looking through microscopes and excited about what they were seeing. However, some students didn't write down their observations, and at the end of the data-collection time, they had nothing to include in their lab reports. We encouraged students who had collected data to share with their classmates in an effort to keep all students moving forward. Many students this year struggled with conclusion writing, much more so than in the first year. Sandy presented the class with sentence starters to get them going in the right direction. On the one hand, this removed a lot of the critical thinking that we were trying to encourage. However, students were able to write their own conclusions on the plant lab, demonstrating that this scaffolding may have been necessary for this particular group of students. On the plant lab, graphing and analysis was done before the lab
LONGITUDINAL EXPERIMENTS IMPROVE UNDERSTANDING OF NATURE OF SCIENCE

report was due and was graded as a separate assignment. More time was spent on these skills this year than last year and the difference showed up in better performance on the analysis and more students turning in the assignment. This lab was tied to class discussions about genetics and students demonstrated understanding of this concept. Overall, in year two, there was a greater level of student participation and buy-in. This engagement was evidenced by a higher rate of students turning in labs, the grades on these labs, and student performance on the unit test.

Discussion

Studies have shown that students do better in science when they are engaged in the classroom (Blumenfeld, 1991). For some students, who don't feel that they have been successful at science, this engagement can be triggered by simple events, such as getting a plant to grow or seeing an organism under a microscope. In this paper, we presented preliminary results of incorporating longitudinal labs into a middle school life science classroom. While these results included statistically significant differences in assessment scores between students involved with the longitudinal labs and a control, the level of student and teacher engagement that these labs inspired are an equally notable contribution of this work. Based on the level of continuity that we observed in the classroom, we believe that every student got something out of these activities. Some advanced students took an interest in graphing and voluntarily produced graphs that required more than the effort needed to just get a good grade. The second year of implementing these labs ran even more smoothly than the first, demonstrating their potential for long-term success.

References


