Evaluating the Benefits of Technology-Enabled Formative Feedback in the Science Classroom

Kimberle Koile, Nathan Kimball, Sarah Pryputniewicz
{kkoile, nkimball, spryputniewicz}@concord.org

The Concord Consortium, 25 Love Lane, Concord, MA 01742
www.concord.org

Paper presented at the 86th NARST Annual International Conference April 6-9, 2013.

This research was supported by the National Science Foundation under grant No. DRL-0733299. Co-Principal-Investigators included Marcia Linn at University of California, Berkeley, and Jim Slotta at University of Toronto. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Abstract

Computer-based activities can provide students with innovative resources such as digital sensors and computational models and can give students opportunities to engage in thoughtful and engaging inquiry, write explanations using the data they have explored, and reflect on the results. When students work with such activities, however, teachers often find it difficult to gauge student understanding, progress, or level of effort expended. The goal of the LOOPS project is to build and evaluate the effectiveness of technology that supplies teachers with timely formative feedback that provides insight into student learning. Timely and effective feedback is widely acknowledged to be a powerful means of improving learning. With today's technological developments, "timely" can truly mean "in real time", right when students are first learning a particular concept. We aim to provide teachers with feedback on students' work, and we have developed a system that delivers that feedback in real time: Student work is summarized, aggregated, and organized to be quickly understood in the fast-paced classroom environment. This paper reports on research on the impact of our real-time formative feedback system on teacher practice and student learning using a motion and graphing curriculum for middle school. In this study, teachers were able to view student work created via activities that employed computational models, motion probes, graphs, and multiple choice and open-response questions. Teachers also were able to view summary information that provided insights into the method and extent of student interaction with the activities. The study compared five sixth grade teachers' use of the technology as they employed real-time formative feedback in two different ways: to hold class discussions focused on examples of their students' work, and to identify and work with students they thought might need help. The goal of the study was to investigate whether varying the balance between these different teaching patterns would result in measureable differences in student learning gains.

Evaluating the Benefits of Technology-Enabled Formative Feedback in the Science Classroom

At the heart of science learning is students' work in collecting, analyzing, and presenting data, and building understanding through classroom conversation and scientific debate. Technology opens up new possibilities for these activities: Students now can collect data via digital sensors, interact with computational models of physical phenomena, share data via wireless networks, and analyze data using computer tools. In a busy classroom of students engaged with computer-based activities, it is difficult for a teacher to know how well her students are investigating and learning, or how far they have progressed through the activities. The NSF-funded LOOPS (Logging Opportunities in Online Programs for Science) project addresses this challenge by developing and researching technology that provides teachers with timely formative feedback. With such feedback, teachers gain insight into student thinking and can modify instruction based on that insight; students can reflect on and revise their thinking. Timely feedback is widely acknowledged to be a powerful means of improving learning, especially when the feedback is formative, i.e., used during instruction for the purposes of guiding learning rather than evaluating learning (Black 1993; Black & William 1998; Duschl 2003). With today's technological developments, "timely" can truly mean "in real time": Technology can provide teachers and students with formative feedback, aka formative assessment, information right when students are first learning a particular concept. It also can enable transient informal assessment information (Bell & Crowe 2001; Ruiz-Primo & Furtak 2007) to persist. The study reported in this paper investigated the use of the LOOPS technology in providing real-time formative feedback in five Boston-area middle school classrooms, investigating the question of how technology can facilitate teachers' use of students' computer-based work to engage and guide students and to influence learning. In particular, the study compared the middle school teachers' use of the technology as they employed real-time formative feedback in two different ways: to hold class discussions focused on examples of their students' work, and to identify and work with students they thought might need help. The goal of the study was to investigate whether varying the balance between these two different teaching patterns would result in measureable differences in student learning gains.

In the sections that follow, we describe the LOOPS classroom culture and curriculum, give specific examples of the technology in use, and detail our methods and results. We then discuss findings, future research directions, and conclusions.

The LOOPS Classroom

In a LOOPS classroom, students, typically in pairs, work through computer-based activities, interacting with computational models and instruments such as motion probes. As they work through the activities, they submit their predictions, data, observations, and reflections to the teacher's machine in real time via a wireless network. They, for example, might submit a hand-drawn prediction graph or a graph created using a motion probe. As the students collect additional data or change or improve upon their ideas, they can resubmit their work. The student submissions constitute the first half of a formative feedback loop. The second half of the loop begins when the teacher receives the student work. The teacher can view the work on a mobile tablet computer while circulating through the classroom observing and interacting with students. Students' text responses, graphs, screen shots of model states, and images created with a drawing tool are arranged for the teacher in tables. Multiple-choice responses are compiled

into histograms. To lead a discussion using student work, a teacher can select some or all student responses for public display: Selected student responses appear, with or without student names, on a class projector and simultaneously on each student's machine. Students become greatly invested in seeing and explaining their data or defending their results. By viewing a representative sample of student results, students whose answers are outliers may decide to rethink their approaches or advocate for their positions. Teachers can use these discussions to redirect student learning if misunderstandings remain or to bring issues to debate or closure.

The amount of information the technology makes available unfortunately can at times be overwhelming for teachers, especially since students are encouraged to reflect and improve on their work and resubmit it to the teacher. To help manage the information, LOOPS provides summary information, such as how many experimental trials each student has completed and how far each student has progressed through the activities. The teacher also can monitor student progress by viewing summary information for the whole class, e.g., a pie chart representing how many students have started and how many have completed particular steps in an activity. She can view a composite of all student graphs for a particular question on a single set of axes and also see how individual student graphs compare with an expected answer. Examples of student work and summary information are shown in the Technology section below.

It is important to point out that the technology in a LOOPS classroom is designed to recede into the background, while the teacher guides students in their role as scientists. With student work and discoveries as the focus of classroom discussions, the class is transformed into an authentic scientific community in which experimenters test and examine the validity of scientific claims, sharing data in order to make sense of it and to examine hypotheses about how things work. These class discussions are opportunities for students, moderated by the teacher, to cast a critical eye on data and ideas. The discussions can serve to reignite experimentation or bring closure to the big topics in the curriculum. LOOPS enhances the classroom experience by enabling these discussions to take place in real time: There is little lag between student experimentation and community examination of results.

Curriculum

One of the goals of the LOOPS curriculum developed for our research is for middle school students to develop a mental model of motion represented as a position vs. time graph. We also want students to understand this graph as a multi-faceted representation of movement in space that could be described both qualitatively and quantitatively. Additional goals are to guide students to view speed not only as an algebraic quantity, but also as a line in an x-y-plane with an associated frame of reference, and to enable students to move from a graphical representation to an algebraic one and vice versa.

Focusing on interpretation of motion graphs first qualitatively and then quantitatively works well with the technical capabilities of the LOOPS technology. Students learn early in their work with the LOOPS system that they may revise their answers to questions while preserving a history all previous submissions. This capability allows for continuous improvement without penalty. So students may answer a graph-reading question such as, "Where is Chico [a dog on a walk] going the fastest? How can you tell?," first qualitatively, i.e., where the position vs. time line on the graph is the steepest, and then quantitatively by calculating the speeds of several of the graph's segments and comparing them numerically.

Our curriculum is not heavily scaffolded with ancillary supports, but rather, the central experience of the curriculum provides several approaches to learning about speed (kinesthetic with

probes, visual with models and graphs, and numerical with measurement and calculation) and then gives the teachers ample access to student work so that they may decide the best course of additional support. Transitioning students from making qualitative arguments to quantitative ones proved to be an excellent vehicle for teachers and students, as evidenced by students' improvement in crafting clear explanations of how they arrived at their answers to questions posed. The crafting of those explanations was intentionally supported well by the technology, so that in addition to learning about motion and graphs, students learned how to construct clear, evidence-based explanations: The technology enabled students to submit graphs linked with open response answers, so that students could create graphs and then use their graphs in fashioning their explanations; and the technology facilitated revision, enabling students to go back easily to any question and revise and resubmit their answers.

A brief overview of the curriculum, organized as four guided inquiry activities, follows. In the first activity (Missing Manual), students explore qualitative aspects of graphing and motion. Using a motion probe students observe the correspondence between the shape of the curve on a position vs. time graph and the kinesthetic motion of their bodies that created it. The learning goals are for students to relate the slope of the line to the speed and direction of their motion and the height of the line to their position in front of the probe.

In the second activity (Modeling Motion), students explore motion by manipulating a model that is connected to a graph, and also the reverse, namely drawing a graph to control a model. This activity begins the transition from a qualitative description of motion to a quantitative one. Here they begin to understand that graphs tell stories and that motion takes place in a frame of reference. They start to relate positions to actual locations and their numeric representation, explaining what a position-time graph represents. Finally, they start quantitatively comparing speeds by seeing how much distance is covered in a given interval of time, and they construct evidence-based arguments to answer questions about speeds on graphs.

In the third activity (Making Measurements), students return to using the motion probe. This activity emphasizes quantitative measurements made from the graphs of their body motion. After practicing making measurements of distance and duration, students use those skills to calculate speed. Students also formally capture their understanding of the frame of reference by drawing the number line defined by the probe, locating a picture of the probe on it.

In the final activity (Telling Stories), students practice interpreting graphs to elucidate the stories being told and drawing graphs from stories. They further practice their ability to turn slopes on graphs into calculated speeds and vice versa.

Technology

The classroom setup is as follows: Students work in pairs using a pen-based tablet computer which also has a keyboard; the teacher circulates carrying a tablet computer with the keyboard folded out of the way; and a tablet (or laptop) computer is attached to a projector. Students interact with computational models , create graphs by hand or using motion probes or models, and provide answers to open response and multiple choice questions. (See Figure 1a.) They submit their work to the teacher's machine in real time via a local wireless network. The teacher can view the student work easily as he carries his computer with him (see Figure 1b) and uses three kinds of information to identify groups who need help and to conduct class discussions based on student work: group progress information, summaries of each group's interaction with models and probes, and each group's submitted work.

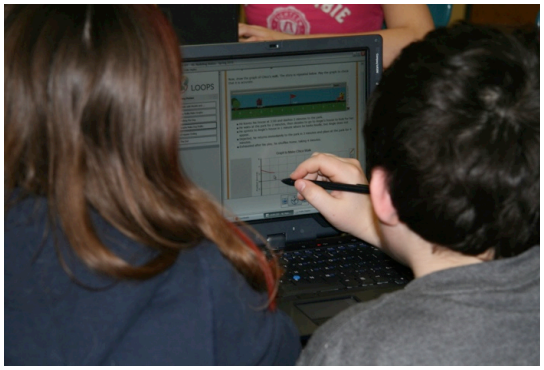
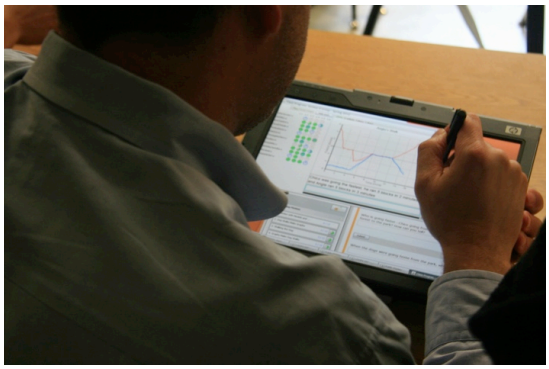


Figure 1a. Students working with LOOPS



1b. Teacher working with LOOPS

Group progress information is provided at two levels. Although these features are very simple, and relate to class management rather than teaching content, they were very popular with and highly used by our field-test teachers. In the teacher's view of an activity (see Figure 2 below), each step in the left-hand navigation panel has a pie chart that shows the proportion of the class that has submitted answers for that step. At a glance, teachers can get an overview of how much work has been submitted. In this example, the pie charts appear on steps 2 through 5. In addition, the numbers to the left of each pie chart show how many student groups are on a particular step, for instance 4 student groups of 12 are working on step 4, and 6 groups of 12 are working on step 5.

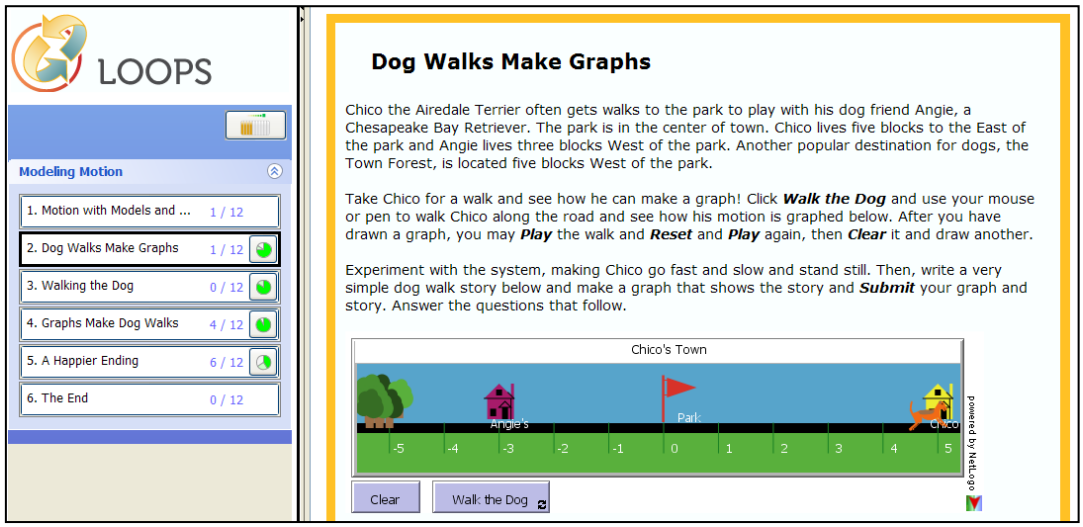


Figure 2. Teacher's view of an activity: left panel is for navigation between steps and overview of student progress, main window shows selected step in the activity

The rectangular button just above the navigation panel brings up the second level of progress information—the Class Progress window, an example of which is shown below. Here, each student group is listed (shown here with anonymized student names), and the step on which each group is working is identified by a blue circle. The green pie charts indicate the proportion of questions answered by each group so far for a particular step. Teachers view individual student answers by clicking on the green pie charts. The particular student work shown is indicated by a red box outline. In the example shown below in Figure 3, the red box indicates that work for the first listed group's question 1.4, i.e., step 4 of activity 1, is being displayed on the right.

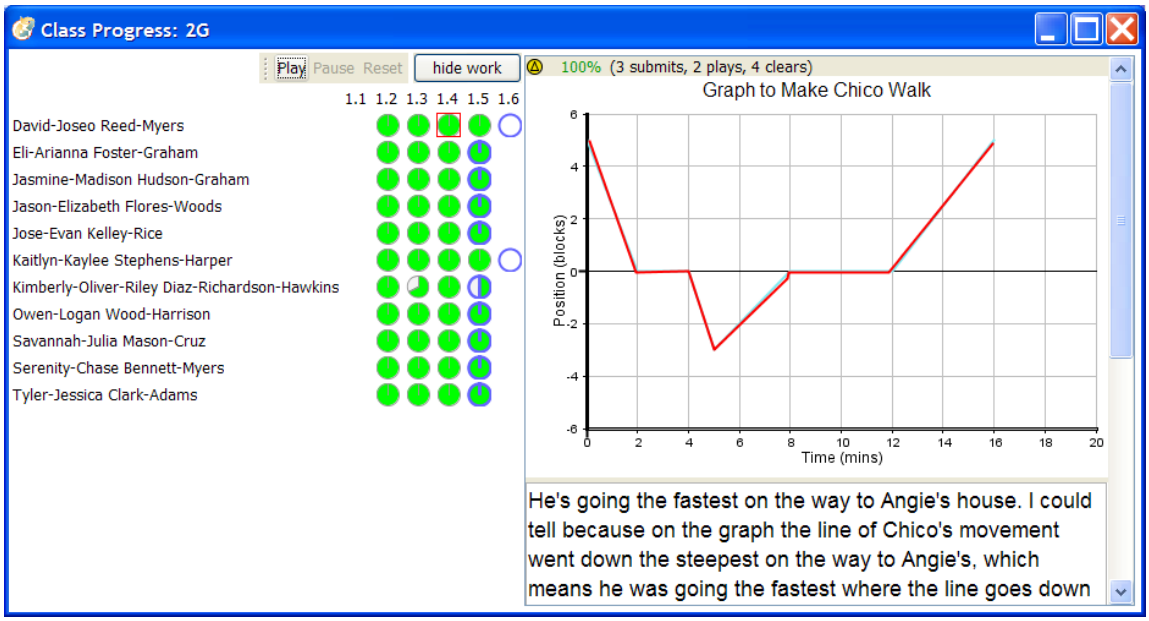


Figure 3. Class progress window with student data displayed on the right; above the student graph is summary information and access to a history of submissions (accessible via the triangle icon)

Summary information that indicates students' level of effort appears in both the Class Progress window, shown above in Figure 3, and in the Student Work window (described below). It appears just above individual student work and includes the number of times each group modified and resubmitted an answer (submits), reran a model (plays), and started over (clears). Teachers also can view a history of each group's submitted work to gain insight into how the group's thinking has changed over time.

Included in the summary information above a student's graph is a computer-generated score of how well the graph matched a teacher-supplied rubric. The score for the graph shown in Figure 3 is 100%. Graph-scoring rubrics are defined in terms of graph segments; the scoring routine uses the rubrics to evaluate segments in student graphs, then returns a total score. The rubric for the problem in Figure 3 is shown in Table 1; evaluated graphs are shown in Figure 4. Students were asked to draw a graph that matched a story, then run a model, shown in Figure 2, to verify that the character in the story, in this case a dog named Chico, moved correctly.

Table 1. Story and scoring rubric based on segment change in x and y values (Δx, Δy) and slope (m)

Story	Segment	Segment Characteristics
• He leaves his house at 3:00 and dashes 2 minutes to the park • He waits at the park for 2 minutes, then decides to go to Angie's house to look for her. • He sprints to Angie's house in 1 minute where he barks loudly, but Angie does not appear. • Dejected, he returns immediately to the park in 3 minutes and plays in the park for 4 minutes. • Exhausted after his play, he shuffles home, taking 4 minutes.	1	Start at $x = 0$, $y = 5$; $\Delta x = 2$, $\Delta y = 5$; $m < 0$
	2	$\Delta x = 2$, $\Delta y = 0$
	3	$\Delta x = 1$, $\Delta y = 3$; $m < 0$
	4	$\Delta x = 3$, $\Delta y = 3$; $m > 0$
	5	$\Delta x = 4$, $\Delta y = 0$
	6	$\Delta x = 4$, $\Delta y = 5$; $m > 0$
	7	Optional segment, $\Delta x > 0$, $\Delta y = 0$

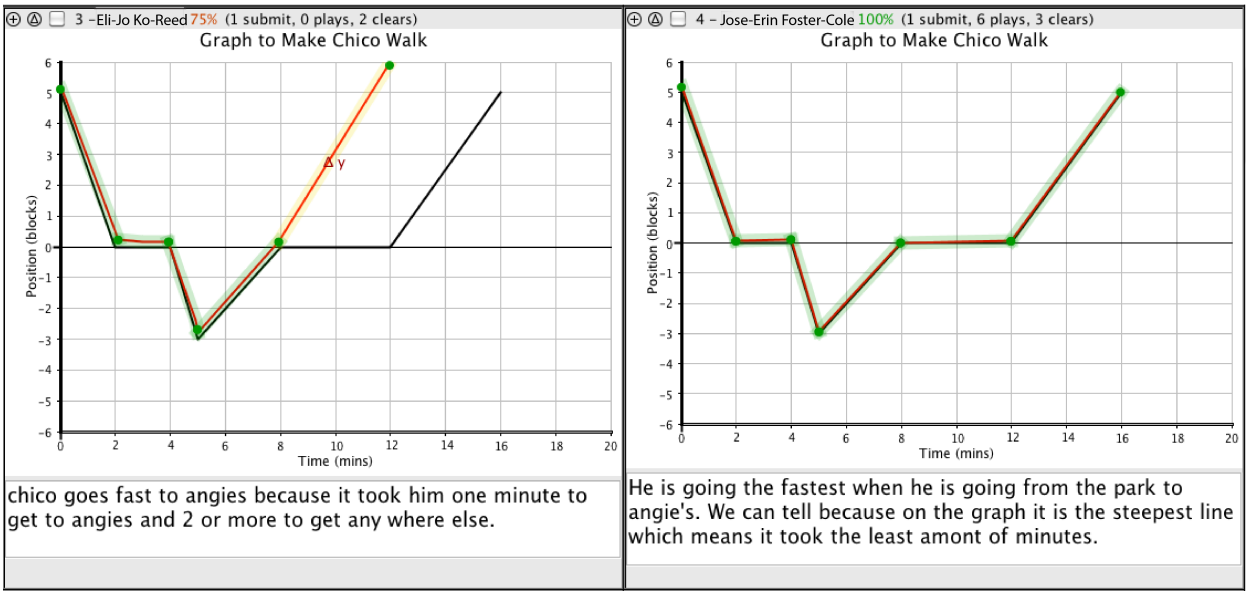


Figure 4. Teachers can view evaluations of student graphs

The Student Work window shows both summary data and a comprehensive display of all submitted student work. As a support for teachers, it also shows sample answers (identified by a cyan box). In Figure 5 below, the sample answer is on the left, and a composite of all student graphs for that problem is on the right. The composite graph image presented here was taken at the end of the field test, after discussions and student revisions, and it shows a convergence on the correct answer. While students work toward convergence, teachers can glance at the composite graphs to identify outliers easily and decide on the best course of action, e.g., either working with individual student groups or initiating a class discussion.

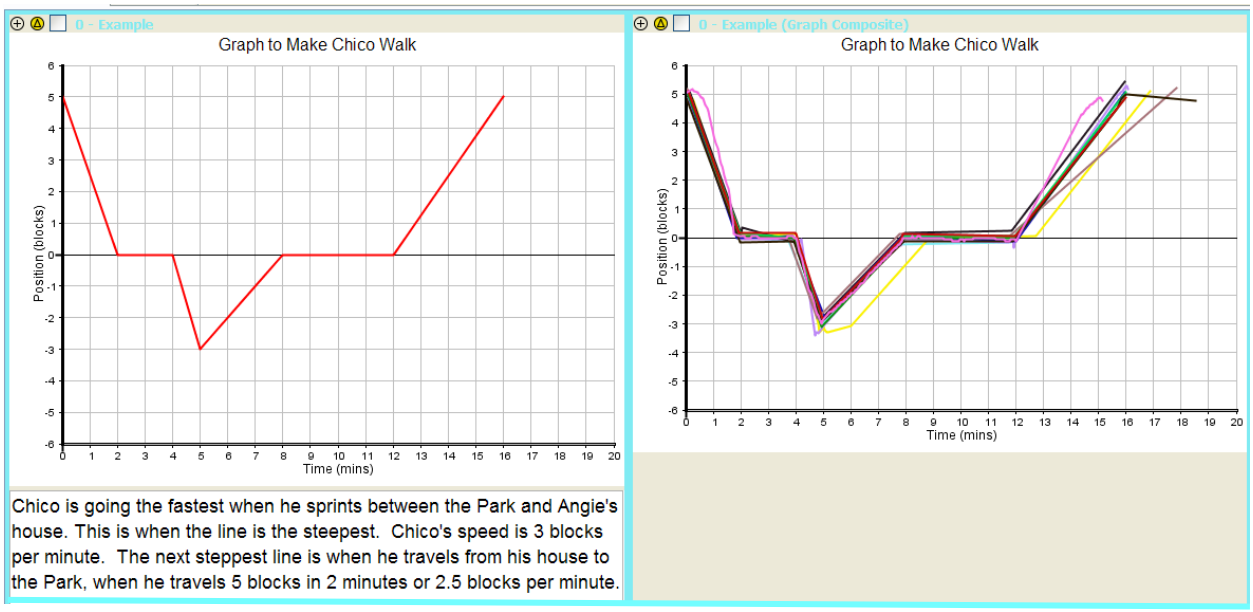


Figure 5. Teachers can view a sample answer (left) and a composite of all student responses (right)

Teachers may project sample answers, composites, or student work using the Student Work window by tapping on the square checkboxes above the entries; they zoom in on an entry using the + icon. (See top of Figures 4 and 5.) When a Project button (not shown) is tapped, the selected entries are transmitted wirelessly to the Public Display computer attached to the projector. In a classroom equipped with a Smartboard, teachers can manipulate the Public Display image using the board.

Methods

Research Design

This research study focused on comparing different teaching patterns for using real-time formative feedback. In particular, we focused on two patterns—using the information for class discussion or for identifying and working with students who needed help. To investigate these patterns, we controlled for when teachers conducted class discussions. In experimental classes, discussions took place at the beginning of class *and* at least one other time during the class. In control classes, discussions took place *only* at the beginning of class, e.g., to discuss previous day's work, and teachers spent the rest of the class working individually with students. The teachers served as their own controls: half of their classes were experimental classes, half were control. In both kinds of classes, the teachers worked individually with students whenever they were not conducting class discussion.

This research design reflects our interest in investigating the use of teaching patterns. We did consider, however, comparing "LOOPing" with "non-LOOPing", i.e., removing for one treatment group teachers' ability to see student work in real time. There are complicating issues, however, in attempting to compare a LOOPS classroom with a non-LOOPS classroom. In particular, the use of probes and models was central to the curriculum, so a non-technology comparison group was not possible as a research design. In addition, we did not have resources for a large enough N to control for teacher effect, so the teachers needed to be their own controls. In such a situation, had a teacher used LOOPS technology in one classroom, but not another, what he or she learned in the LOOPS classroom would likely have influenced the teaching in the non-LOOPS classroom. Therefore, a research design in which teachers were asked to do without the real-time information in some classes was not a practical one.

Participants

Participants included five sixth-grade teachers in three Boston-area schools and their 392 students. Three of the teachers had used the LOOPS technology and curriculum the year prior to this study (Teachers A, M, and K, see below); two teachers had no prior experience with either the technology or the curriculum (Teachers C and S). The teachers' years of classroom teaching experience ranged from three to more than 30. A summary of participants is shown below in Table 2.

Table 2. Summary of teacher participants and their schools

Teachers	Number of Students	School Town	SES free or reduced lunch	School's Student Ethnicity (% non white)
M, K, S	186	Newton	16%	7% AA, 13% AS, 8% L, 8% Multi (36%)
A	112	Arlington	12%	4% AA, 10% AS, 5% L, 2% Multi (21%)
C	94	Reading	3%	2% AA, 3% AS, 1% L, 1% Multi (6%)

Professional Development

In preparation for each field test run, researchers met with teachers to review the curriculum, technology, and experimental design. For each school, we met with field-test teachers for two or three one-hour sessions. This preparation was barely adequate, but the meeting times were difficult to fit into teachers' schedules. As mentioned earlier, three of our five field-test teachers were veterans of LOOPS and had become comfortable with the curriculum and technology. They had been involved in design and testing phases, and one had helped us redesign parts of the curriculum based on her experience with early versions of the technology and curriculum. (For our collaborators' study investigating the role that professional development can play in using embedded formative assessment information to aid teachers in improving inquiry teaching and learning see (Gerard, Spitulnik, & Linn 2010).)

The three veteran teachers were largely unassisted during the field test. The two new LOOPS teachers received help during the field trial, as the veteran teachers had in previous years. This help included researchers offering suggestions and explanations about instructional options, e.g., which student responses might foster effective class discussions. This approach enabled the new teachers to feel at ease with the research, curriculum, and technology.

As an aid to field-test teachers, we provided both a software guide and a curriculum and field-test experiment guide. The software guide is a "how-to" guide for our LOOPS portal and software features. The curriculum guide, lists for each step of the curriculum what the learning goals are, and anticipated problems that students may have, based on previous year's work. Teachers did report using these guides when preparing individually for the field tests.

Data Sources

Data consisted of individual student pre- and post-tests that included multiple choice questions, open response explanation questions, and a question asking students to draw a graph that corresponded to a motion story. Other data included the computer-logged history of all work submitted by students and logs of teachers' projected public displays of student work. This student work, which we considered to be embedded assessments, included the same kinds of questions as on the pre- and post-tests, as well as questions asking students to create graphs using motion probes, write their own stories and create graphs to match, and take measurements to calculate speed. In addition, students crafted explanations using data and graphs to support their calculations and reasoning.

In addition to the computer-logged data, we took observation notes and videotaped all class discussions. These data enabled us to investigate patterns of students' submissions and teacher discussions: We constructed timelines of student submissions for each class in order to identify student work created before and after class discussions and to evaluate the effect of class discussions on the quality and rates of revision of students' work. The timelines also enabled us to characterize teacher practice by providing a way to see patterns of classroom interaction.

Finally, we collected data from interviews and meetings with teachers and from a teacher workshop this past summer. The interviews were conducted, transcribed, and codified by our project evaluator.

Analysis and Results

Teacher Practice

The objective of our LOOPS technology development has been to solve the problem of providing teachers with immediate, helpful formative feedback that they can readily use for effective teaching. It was our intent that the technology be a tool, adaptable to a wide range of teaching styles in the context of the guided inquiry activities that form the basis of this study. We, thus, did not want our system to constrain teachers to a particular teaching style or method. Therefore, in our field trials and subsequent analysis of field-test data we have attempted to characterize the teaching styles of our five field-test teachers, their appropriation of the technology, and their approach to using it. We describe here the five use cases illustrated by our teachers. The variation in our teachers' use of the LOOPS technology provides valuable information about the technology's generality and potential for broader application.

If necessary, teachers were reminded by researchers of the need to conduct a class discussion (or not) during classroom activities according to the research protocol in order to maintain fidelity of the field-test implementation. (See (Furtak et al. 2008) for an investigation of implementation fidelity for embedded formative assessments.) In addition, during each field trial, researchers discussed with the teacher the classes' on-going progress and plans for the next day. These conversations occurred after class, after school, or over email as need required or time allowed.

Teacher Characteristics. The five field-test teachers, identified by letter, were introduced in the Participants section above. This section offers further characterization of their experience as teachers, their teaching styles, and their use of LOOPS curriculum and technology.

Teacher A: Experienced teacher (30+ years) certified in multiple subjects (science, health, language arts) who has told us stories that indicate initiative and creativity in bringing technology into the classroom on a very limited budget. Teacher A had participated in two previous studies. Students in A's classes did not have strong math skills, which was reflected in their pre-post and embedded assessment scores. A's interaction with students during class discussion was Socratic—never revealing an answer, but eliciting answers from students. As our most experienced veteran of earlier LOOPS field trials, A used the teacher tablet effectively, looking at student submissions while circulating among students during their independent work, checking that answers met expectations. A was able to select student submissions for discussion and think on-the-fly about discussion points. When leading a class discussion, A demanded student attention (and got it) and was reluctant to move on until perceiving that each student understood the current discussion topic.

Teacher M: Experienced teacher (15+ years), originally a math teacher, but now splits time equally between math and science, teaching both subjects to the same students. M would look at student work overnight and select examples for the next day's opening discussion. In class, M would ask the students focused questions during discussion. M was less comfortable with real-time discussions but appeared to become more comfortable with time. M often gave students 30 to 60 second opportunities to discuss questions with their partners before asking students to speak, which seemed to improve the quality of student answers. In a marked change from the previous year's experience, M made regular use of the teacher tablet in this stud and used the submitted student data to identify student groups who needed help.

Teacher K: Experienced (10+ years) and splits time between math and science, teaching both subjects to the same students. Discussion style was the most "instructional" of our teachers, teaching concepts, e.g., speed, as opposed to eliciting student ideas to guide a discussion. K ranked among the top of our teachers in the proportion of time spent projecting and discussing steps of an activity, as opposed to using student work as the focus of discussions. K seemed to favor a discussion style often used in teaching math—working a problem through with the whole class. K was the most brief of our teachers, spending the smallest percentage of class time in class discussion. K was quite facile with technology and used the teacher tablet to assess class progress and identify and assist struggling students. K was adept with Smartboard technology and used it to mark up graphs and carry out calculations on the board. K often had students come to the front of the classroom to annotate what was being projected on the Smartboard.

Teacher S: Novice (3 years) teacher who was new to the school district. S frequently used students' work in class discussions, usually starting discussions by displaying work and asking students what they noticed about it. Students appeared familiar and comfortable with this style of discussion and were quick to offer ideas. Teacher S was skillful in eliciting student ideas and used positive encouragement while being insistent on clarity. Discussions were lengthy and seemed productive in gaining consensus about how to solve problems. S often reinforced the need for justification in explanations. S was new to LOOPS and did not often carry the teacher tablet or view student submissions while students were working. S preferred to circulate around the classroom and look over students' shoulders to see how they were doing. S did select student work for discussion, often getting assistance from researchers, as was consistent with our professional development protocol for first-year LOOPS teachers.

Teacher C: Experienced (15+ years), with many years spent teaching health and biological sciences. C was new to LOOPS. C was less experienced in teaching physical science and did not seem comfortable with the subject matter. This unease with content contributed to C's nervousness with free-flowing discussions of graphs and motion. This nervousness, however, did not prevent C from fully using the LOOPS technology. C carried the teacher tablet during class and examined student submissions while circulating among students during independent work time. C was eager to use student work for discussion, and researchers helped C to select student work for discussion, consistent with our professional development protocol. In class discussions, C tended to use an in-depth question and answer session with individual students rather than soliciting ideas from the larger group. This technique probably contributed to the low attention span we observed for many of C's students during class discussions. C's students seemed reticent to offer their ideas in the larger discussions.

Using Real-Time Information. One important question from our LOOPS work is: Can teachers cope with all the information generated in the course of an active inquiry-based classroom? We believe that the answer to that question is yes, especially after having experience with the curriculum and technology. Our evidence for this statement is that all teachers except S carried the teacher tablet during the field trials and viewed student data in real time, while moving about the class helping students. These teachers used the real-time data to identify students who had difficulty with the concepts. All teachers, including S, used the technology to select and project submitted student work for class discussion.

Class sizes in our Massachusetts classrooms were on average 24 students. It is reasonable to wonder if student data from larger classes would have overwhelmed teachers. When we asked this question of the five LOOPS teachers during our summer teacher workshop, all five teachers responded that the technology would be even more important in a larger class because it would enable them to efficiently and effectively identify and assist all of their struggling students; in a large class, they felt that without the real-time technology, they would not be able to identify quickly students who needed help.

Classroom Variation. With each of our five field trials, we attempted to keep conditions as similar as possible by adhering to our research protocol. As expected, variations occurred that illustrate differences in the classes, students, and choices made by teachers. Some of these differences are described here.

For each field trial, LOOPS work was done on seven consecutive class periods with never more than one intervening non-LOOPS day. The pre- and post-tests were conducted on days directly preceding and following the class work. Single class periods were nominally 50 minutes. The trials of Teachers M and K, which occurred simultaneously as their class times did not conflict, was interrupted by one day for a grade-wide field trip; Teacher C's classes met every other day in double blocks, i.e., in classes that were twice the usual length.

To see larger patterns of classroom interaction, we mapped the span of each class's field trial work into a single timeline for each class. Onto these timelines we plotted the time and duration with shadings to indicate the characteristics of classroom discussions. The graph below in Figure 6 gives an overview of the interplay between students working independently and teachers conducting discussions. All teachers' classes are shown, with letters for each teacher (C, S, M, K, and A) and numerical designations for their four (or, in the case of A, five) classes. Gray vertical bars represent discussion using student work, salmon bars represent discussion about an activity (without using student work); hash marks represent discussions during class, no hash marks represent discussions at the beginning of classes. Note that only experimental classes contain discussions during class. Teachers were not constrained as to the length of classroom discussion or the choice of materials to project. We did encourage the use of the real-time student data, but if a teacher preferred to use lessons on the board or a projection of the activity, these actions were acceptable. Some teachers almost always preferred to use student data (gray-colored bars in Figure 6) during discussions, while others preferred teaching using the activity pages (salmon-colored bars in Figure 6) rather than student work as the focus of discussions.

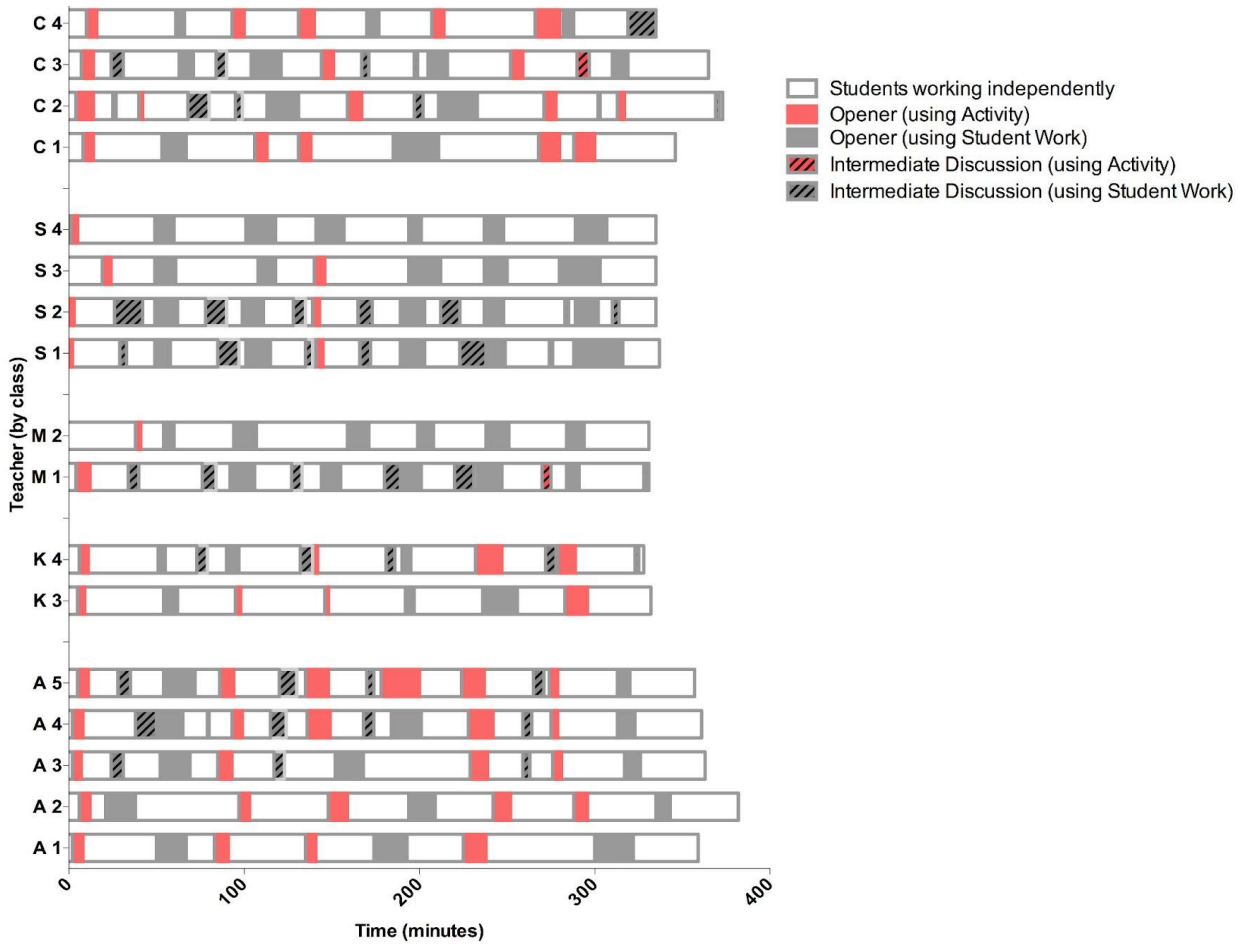


Figure 6. Distribution and content focus of class discussions for each teacher for each trial

The graph in Figure 7 compares the teachers' total time spent in class discussions. Teachers in experimental classes always spent more time in class discussion than control classes, as would be expected if the field trials were properly conducted. Experimental classes averaged over 30% of class time devoted to discussion; control classes averaged just over 22% of class time spent in discussion. The variation between teachers, however, is marked. Comparing control classes, the variation spans from 17% of class time (Teacher K) to 26% of class time (Teacher S), and in experimental classes, total duration of discussion times ranged from 24% of class time (Teacher K) to nearly 40% of class time (Teacher S).

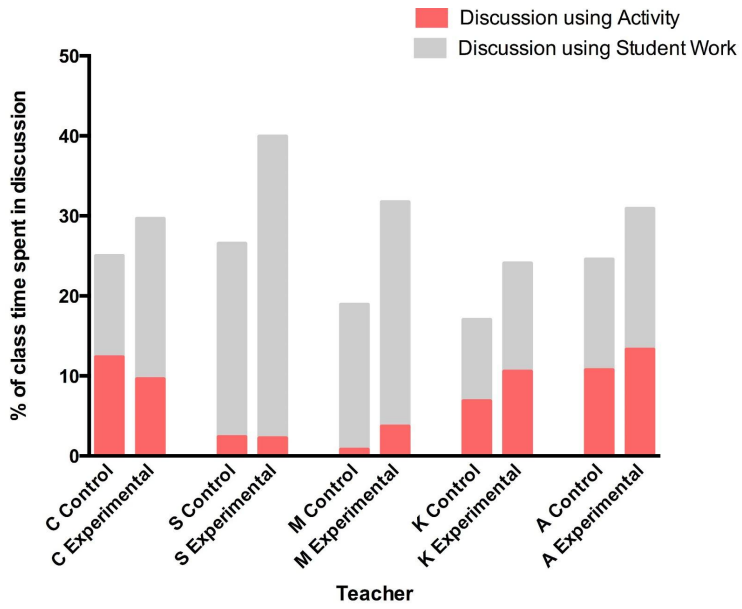


Figure 7. Graph comparing the percentage of time each teacher spent in class discussion during their trial

Examining the percentage of discussion time spent using student work vs using an activity page as the basis of discussion reveals another significant difference in teaching style. Teachers M and S, for example, greatly preferred using student data in discussions, as shown in Figure 8.

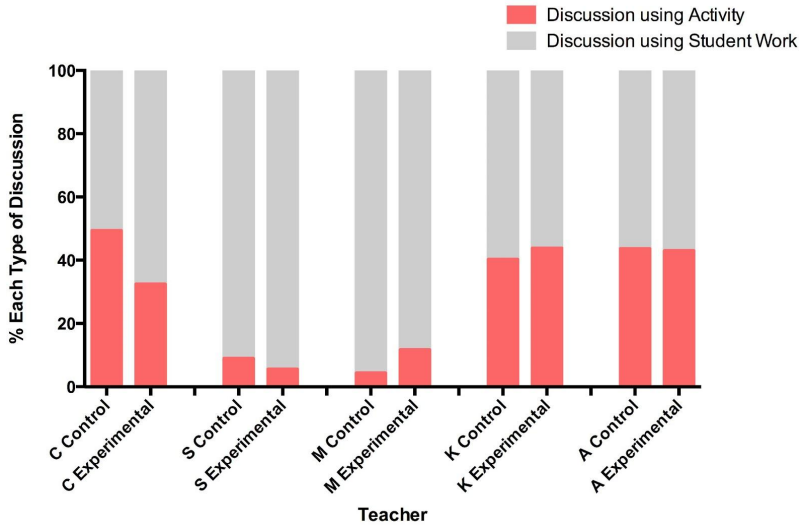


Figure 8. Graph comparing the percentage of time each teacher spent discussing student work vs discussing an activity (without mention of student work)

Given the differences seen among our teachers, it is clear that the LOOPS system was able to accommodate a wide variety of teaching styles. Teachers used their class time in the way they believed would best serve their students. The teachers who spent the least time in discussion (Teachers M and K) believed that their students most needed individual time to work through problems. Their students had fairly strong math skills, and they wanted their students to work individually practicing graphing, measurement, and writing clear explanations. Teacher A had students who struggled with math, so it took

more discussion time for her students to be informed enough to work independently. Teacher S based her teaching style on consensus building and idea generation, so a great percentage of time was devoted to discussion. Teacher C used a discussion style that involved longer exchanges with individuals, which accounted for longer discussions.

One clear outcome of our LOOPS trials is that teachers can and do use real-time formative data. Teachers have flexibility in how they use the data, with respect to their knowledge of their students and the teacher's own individual teaching styles.

Teacher Interviews

Teacher comments from interviews help document their perceptions of the project.

"LOOPing"

Selecting examples helped students stay engaged,

> "I know when students saw an example that was pretty close to their work, they paid a lot closer attention. They need to see that others are doing something similar, that others are making the same mistakes, that they're not alone; it keeps them interested."

Another teacher commented,

> "I just think their working on the computers was very engaging, they were paying attention much better than in other classes, even after the 6th day. I think just knowing their work will be presented and talked about. I think that really kept them interested, it's like a group responsibility."

A teacher commented on how students might perceive submitting their work differently as a result of LOOPS,

> "They see their work almost as soon as they submit it and we can talk about it in the class. It seems to change the idea of what it means to submit work, where before they think it's only about getting a mark the next day, but here they begin to see it's really to have a class discussion."

Diagnostic use

Teachers commented on the value of identifying problems in content understanding,

> "They seemed to be doing what they're supposed to be doing. It was a lot easier for them to submit their work; they simply didn't have to get up and move, and I got everything in a very convenient way. It improved efficiency of my teaching; I was able to discover who had a problem much quicker than otherwise."

Another offered,

> "I think [it was] great as a walk-around diagnostic tool; it helped me understand which students needed help. It's as if I could see the whole class at once and know [who was] getting it and who wasn't, then go over to those students and help them."

One teacher said,

> "LOOPing [provided] great information for me as a teacher; I could focus on those who were having problems. It makes [teaching] more efficient."

And another said,

> "I like being able to know who needs help without making that information public. Students don't need to be embarrassed; I can just walk around the room and talk with students without everyone knowing who is having trouble."

Selecting Examples

Teachers became more familiar with selecting appropriate samples to display to facilitate classroom discussions. While there is no rule or consensus, it seems that most teachers select a good "correct" sample and show it alongside two or more incorrect samples to provide contrast. One teacher found it very easy to select examples, often using friendly competition to motivate students,

> "It's really easy [to choose student work], it's not a problem… I don't always try to pick ones [from the best students] because their [work] is always chosen when we have a project. [If] Johnny knows the answer, I try to not always pick Johnny's because I want to get more people to see if they can figure it out. So if I have somebody else aside from Johnny who can show me a correct [answer], I'm going to [choose] that [one] because that person's going to get the validation, and think, "Oh, it was right, I beat Johnny."

Inquiry-based instruction

The process of revision and working together on problems outlined in the curriculum helped students study and think about science in an inquiry-based manner, on which teachers commented,

> "I like how they really had to explain what the graphs were doing and why, and then work with a partner to gain consensus."

Student Learning

Pre-Post. Students took an individual pre-test before starting the activities, then after the final activity, they took an individual post-test. Students' answers for the pre- and post-tests were scored using graph analysis rubrics, knowledge integration (KI) rubrics (Linn & Eylon 2011; Slotta & Linn 2000) or multiple choice scoring, depending on question type.

Analyses of pre-post tests were done using only complete sets of data. Only when a student completed the same item on the pre-test and on the post-test was the student's answer included in the larger analysis. This constraint was used to ensure that the same population was measured on both the pre-test and post-test.

The pre-post tests contained 13 items: 10 multiple choice, two open response, and one graph-drawing item. The first four items of the pre-post test (three multiple choice items and one open response item) were taken from a pre-post test developed by our University of California, Berkeley collaborators. Other items were developed by The Concord Consortium staff.

All Students. We analyzed student pre-post learning gains collectively and found that students showed significant gains on all 13 pre-post items: multiple choice ($N=291$ to 320; $p<0.003$ to $p<0.0001$), open response ($N=319$; $p<0.0001$), and graph-drawing ($N=294$; $p<0.0001$). One open response item (item 4) and the graph drawing item (item 12) showed significant effect sizes (Cohen's d of 0.67 SD and 1.71 SD, respectively). Effect sizes for multiple choice items ranged from 0.11 SD to 0.41 SD. Data for five of the 13 pre-post items are shown in Table 3 below. The five items include the four items validated by University of California, Berkeley, (items 1-4); and one graph-drawing item (item 12).

Table 3. Pre-post test data for five of the 13 items for all students

	All Classes				
	N	Average	Std. Dev.	Student's t-test (paired, two-tailed)	Cohen's d (SD)
1 Pre	319	0.77	0.42		
1 Post	319	0.87	0.34	< 0.0001	0.16
2 Pre	319	0.24	0.43		
2 Post	319	0.35	0.48	< 0.0001	0.17
3 Pre	319	0.57	0.50		
3 Post	319	0.74	0.44	< 0.0001	**0.26**
4 Pre	319	2.73	0.91		
4 Post	319	3.41	1.12	< 0.0001	**0.67**
12 Pre	294	16.80	7.40		
12 Post	294	21.12	5.32	< 0.0001	**1.71**

Differences Across Control and Experimental. We also analyzed student pre-post learning gains by comparing each teacher's control and experimental groups. Control and experimental groups showed significant gains across all five teachers on all 13 items. Data for the five pre-post items above, separated into control and experimental groups, are shown in Table 4 below. Bolding in the table shows significant effect sizes (Cohen's d: <0.2=small effect size; <0.6=medium effect size; <0.8=large effect size). Highlights show which items had statistically-significant changes from pre to post: Yellow indicates $p<0.0001$ (****); orange indicates $p<0.001$ (***); purple indicates $p<0.01$ (**); blue indicates $p<0.05$ (*). For multiple choice items (1-3), the Wilcoxon signed-rank t-test was used; for open response and graphing items (4 and 12), a paired Student's t-test was used.

In addition, to determine the comparability of the control and experimental groups, we compared the pre-test data for each of the groups. Mann-Whitney U tests were performed for multiple choice items and unpaired Student's t-tests were done for open response and graphing items. These results, a subset of which are shown in the rightmost column (All Teachers: Control Pre vs. Experimental Pre) in Table 4 below, indicate that we are unable to compare learning gain differences between the control and experimental groups because the two populations were very different before using the LOOPS technology and curriculum. We aimed to make the groups as similar as possible in our research design, but probably due to subtle tracking in the schools in which we worked, there were significant differences in the performance of control and experimental groups on the pre-test. The data for items 1-3, 4, and 12, are shown below in Table 4 and in the graphs in Figure 1 in the Appendix.

Table 4. Pre-post test data for five of the 13 items for control and experimental groups

	All Control					All Experimental					*All Teachers: Control Pre vs. Experimental Pre*
	N	Average	Std. Dev.	Paired, two-tailed Student's t-test	Cohen's d (SD)	N	Average	Std. Dev.	Paired, two-tailed Student's t-test	Cohen's d (SD)	
1 Pre	153	0.79	0.41			166	0.75	0.43			
1 Post	153	0.88	0.33	< 0.0001	0.140	166	0.86	0.35	< 0.0001	0.174	*0.4228*
2 Pre	153	0.29	0.46			166	0.19	0.39			
2 Post	153	0.35	0.48	0.0003	0.076	166	0.36	0.48	< 0.0001	**0.256**	*0.0248*
3 Pre	153	0.64	0.48			166	0.50	0.50			
3 Post	153	0.79	0.41	< 0.0001	**0.225**	166	0.70	0.46	< 0.0001	**0.287**	*0.0116*
4 Pre	153	2.90	0.97			166	2.58	0.83			
4 Post	153	3.58	1.08	< 0.0001	**0.678**	166	3.26	1.14	< 0.0001	**0.679**	*0.0022*
12 Pre	136	17.91	6.99			158	15.84	7.63			
12 Post	136	21.29	5.34	< 0.0001	**1.362**	158	20.97	5.32	< 0.0001	**2.015**	*0.0165*

Differences Across Teachers. We looked at all items across all teachers for control and experimental groups, and collectively by teacher. Data for one of the items—the graphing item 12—for experimental and control groups are shown in Figure 9 and in Table 1 in the Appendix. Again, the data show learning gains across all classes for both control and experimental groups. In addition to illustrating the learning gains, the data also show the significant differences in student populations between teachers. In particular, Teacher A's students scored significantly lower on the pre-test on this item than the other teachers' students. Teacher A's students, however, showed significant learning gains. Figure 9 illustrates these differences and gains.

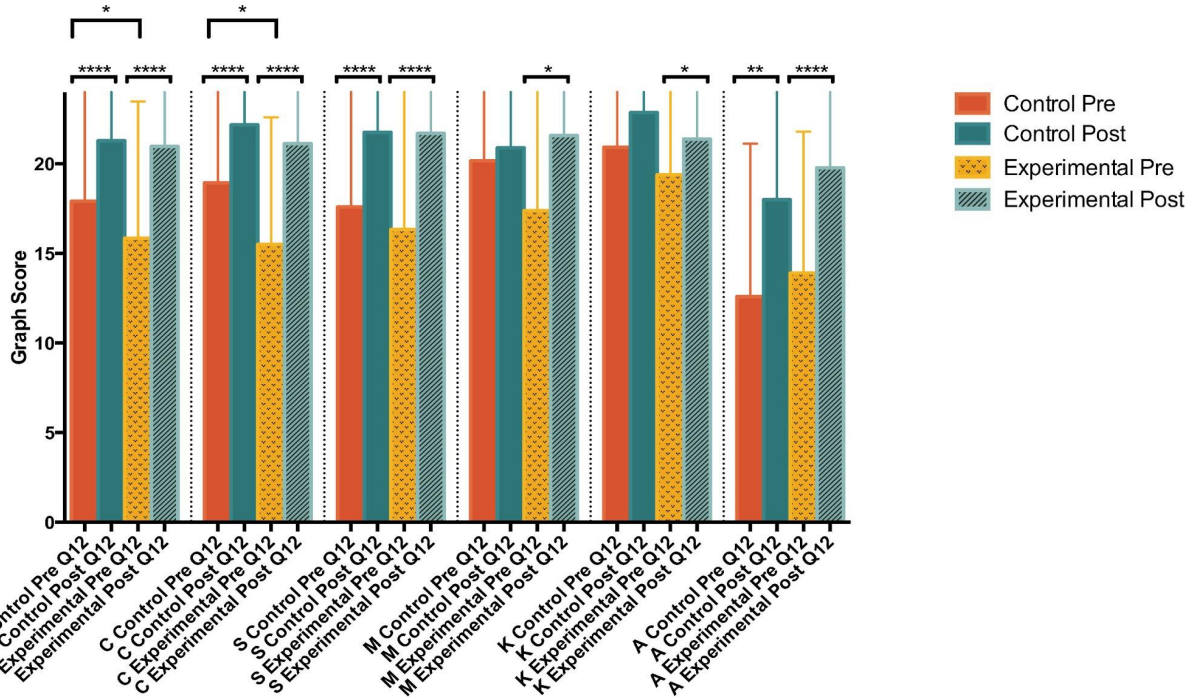


Figure 9. Pre-post data for control and experimental groups by teacher for graphing item 12; $p<0.0001$ (****); $p<0.001$ (***); $p<0.01$ (**); $p<0.05$ (*)

The difference between control and experimental student populations is further illustrated by Table 5, which combines control and experimental groups for each teacher; and the graph in Figure 10. Again note that Teacher A's students' average pre-test score is lower than other students' pre-test scores.

Table 5. Pre-post data showing learning gains for all teachers' students for the graphing item (12)

	All Classes					C Classes				
	N	Average	Std. Dev.	Student's t-test (paired, two-tailed)	Cohen's d (SD)	N	Average	Std. Dev.	Student's t-test (paired, two-tailed)	Cohen's d (SD)
12 Pre	294	16.80	7.40			86	17.21	6.60		
12 Post	294	21.12	5.32	< 0.0001	1.71	86	21.69	4.21	< 0.0001	1.93

	S Classes					M Classes				
	N	Average	Std. Dev.	Student's t-test (paired,	Cohen's d (SD)	N	Average	Std. Dev.	Student's t-test (paired,	Cohen's d (SD)
12 Pre	73	16.97	7.64			39	18.67	7.05		
12 Post	73	21.73	4.78	< 0.0001	1.91	39	21.26	6.07	0.0295	1.01

	K Classes					A Classes				
	N	Average	Std. Dev.	Student's t-test (paired,	Cohen's d (SD)	N	Average	Std. Dev.	Student's t-test (paired,	Cohen's d (SD)
12 Pre	30	20.10	5.33			66	13.47	8.07		
12 Post	30	22.07	4.52	0.0094	0.89	66	19.18	6.69	< 0.0001	2.10

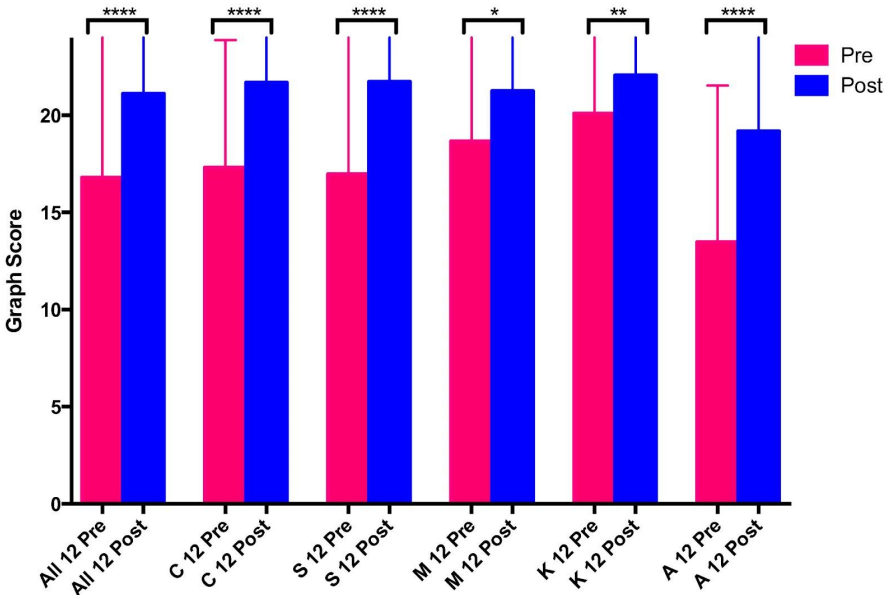


Figure 10. Alternate view of the data in Table 5; data show learning gains for all teachers' classes; $p<0.0001$ (****); $p<0.001$ (***); $p<0.01$ (**); $p<0.05$ (*).

The differences in student performance across teachers, shown for example in the above tables and graphs, illustrate the difficulty in attempting to control for teacher: With small numbers of students, as in our study, teachers need to serve as their own controls because the differences in their student populations may be significant. The results across all teachers, however, are encouraging: Regardless of what students started out knowing, they were all able to make significant gains in learning.

Embedded Assessments. Given our interest in investigating the effect of class discussion on student learning, a set of embedded assessments was selected for analysis. The set included items that asked students to draw a graph based on a given story and answer open response questions about the speed of an actor in the story. The items are identified by the labels *Where*, *Who*, and *When: Where* was Chico [a dog] going fastest? *Who* [of two dogs] was going faster? *When* the dogs were going home, who was going fastest? The items were chosen for analysis because they were representative of the questions throughout the curriculum and pre-post tests, and they could be objectively and reliably scored. The graphs were scored for accuracy against a defined rubric, both by graph-scoring software and by a researcher, and the open response items were scored against a KI rubric. All student submissions were scored, allowing us to see students' learning progress across the lesson.

Student scores were graphed against the timing of class discussions, which allowed us to investigate the patterns of class discussion and the timing and quality of student responses. Shown below in Figure 11 is an example of one such graph for one of teacher S's control classes. The graph enables us to get an overall view of what was taking place in the classroom, e.g., what the teacher was discussing (activity or student work), when students submitted responses with respect to the discussions, and how the responses changed over time. In each graph, vertical bands indicate extent of a class discussion; gray represents discussion of student work, salmon represents discussion of the activity. Each data point is a student submission, and subsequent submissions from same student group are connected with a segment.

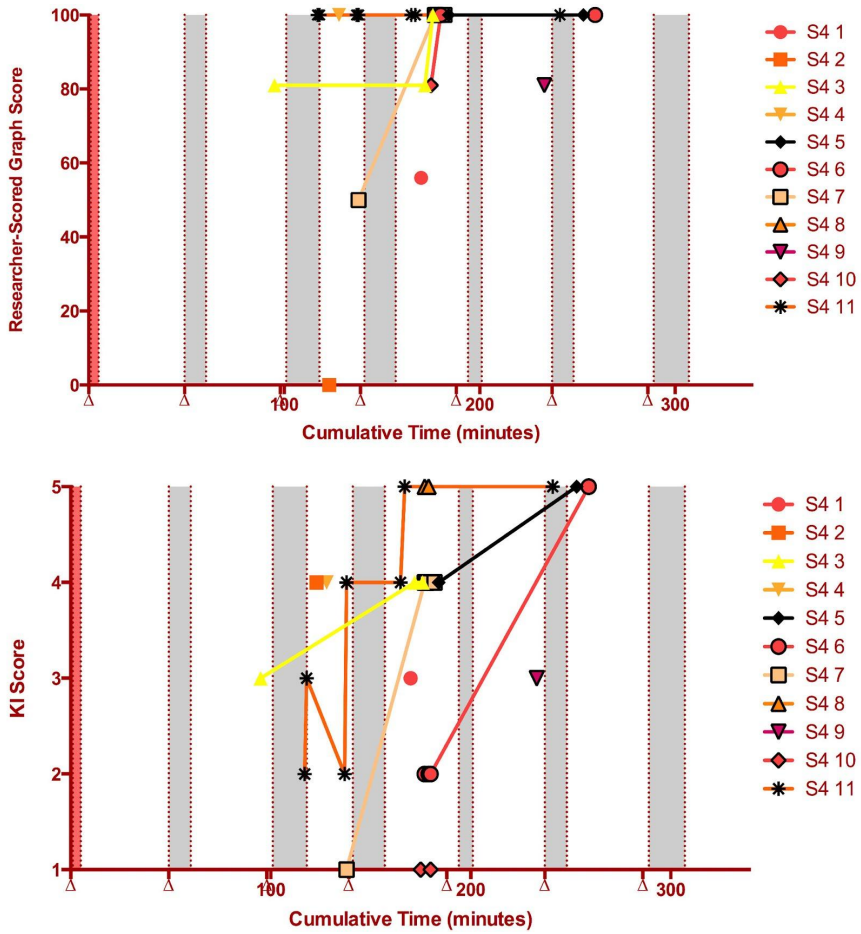


Figure 11. Class discussions and *Who* submissions for one of S's control classes; top shows data for graph question, bottom shows data for KI open response question

Shown below in Figure 12 is an example of a graph from one of A's experimental classes. Note the difference in number and focus of discussions (activity or student work), and the difference in submission patterns for the students. We are continuing to analyze the data using this novel graphical representation.

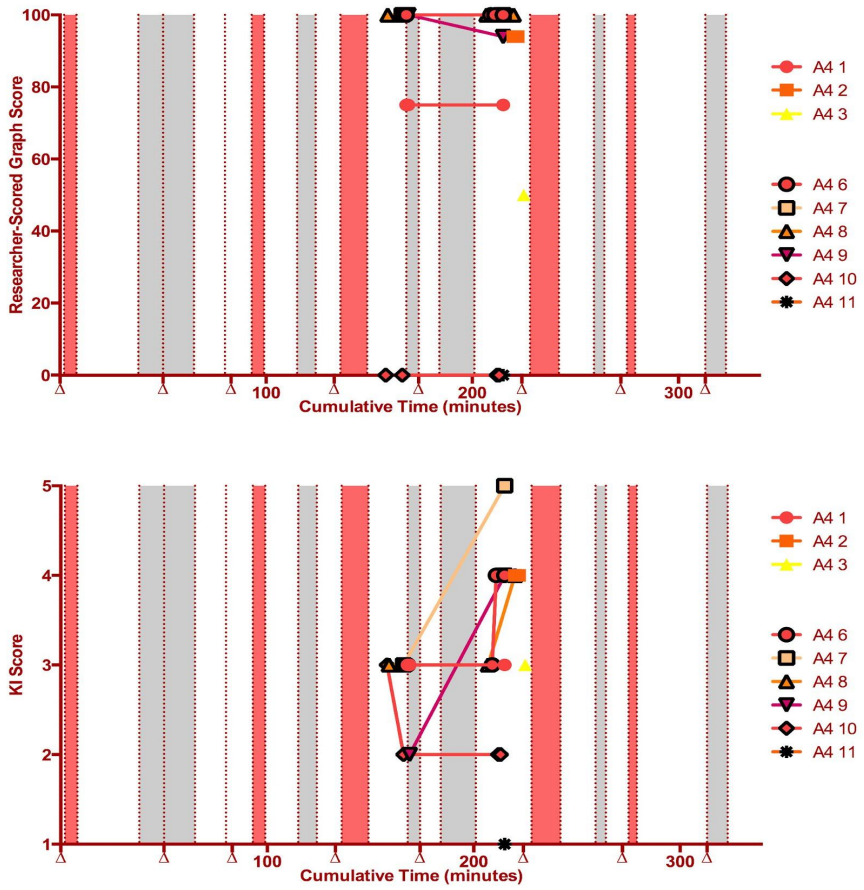


Figure 12. Class discussions and *Who* submissions for one of A's experimental classes; top shows data for graph question, bottom shows data for KI open response question

Revision Rates. We computed the student revision rate for each of the three assessment items discussed above, identified with the labels *Where*, *Who*, and *When*. By revision rate we mean the percentage of student pairs who submitted more than once out of the total number of student pairs. We excluded those student pairs who did not make a first submission until after a relevant class discussion; N = 108, 116, 79 for *Where*, *Who*, *When*; with N = 3, 8, 5 excluded for *Where*, *Who*, *When*, respectively.

We found a revision rate of between 57-72% depending on the question. Between the first and last submission for pairs who revised, student scores increased significantly on all items. On average those who resubmitted had a first answer lower than those who submitted only once, but their last answer scored higher. This finding seems to indicate that re-submitters needed the revision, as evidenced by a low score and large standard deviation on first submission, and the subsequent improved score. So students benefited from revision. Teachers' use of the student submissions in real-time fostered that revision, contributing to a culture of revision in the classroom. The data shown Table 6 indicate that for

all students, the *Where*, *Who*, and *When* embedded assessment items had revision rates of 60.45%, 71.60%, and 58.39%, respectively, for students who had more than one submission.

Table 6. Revision rate data for KI open response questions *Where, Who, and When* items

	Submission	N	Average	St. Dev.	Cohen's d	Student's t-test (paired, two-tailed)	Revision Rate
	First (revisions)		2.60	0.61			
	Last (revisions)	107	3.09	0.81	0.17	< 0.0001	60.45%
Where KI	Single	70	2.79	0.59			
	First (revisions)		2.88	0.86			
	Last (revisions)	116	3.67	0.92	**0.22**	< 0.0001	71.60%
Who KI	Single	46	3.20	0.98			
	First (revisions)		2.66	0.90			
	Last (revisions)	80	3.50	1.06	**0.21**	< 0.0001	58.39%
When KI	Single	57	3.19	0.97			

Below are tables and graphs that illustrate overall revision rates and revision rates for each teacher's students for the KI portion of the *Where*, *Who*, *When* items. Differences in teachers' populations are visible in both the scores and the spread (SD). For each graph, p values are indicated as p<0.0001 (****), p<0.001 (***), p<0.01 (**), p<0.05 (*).

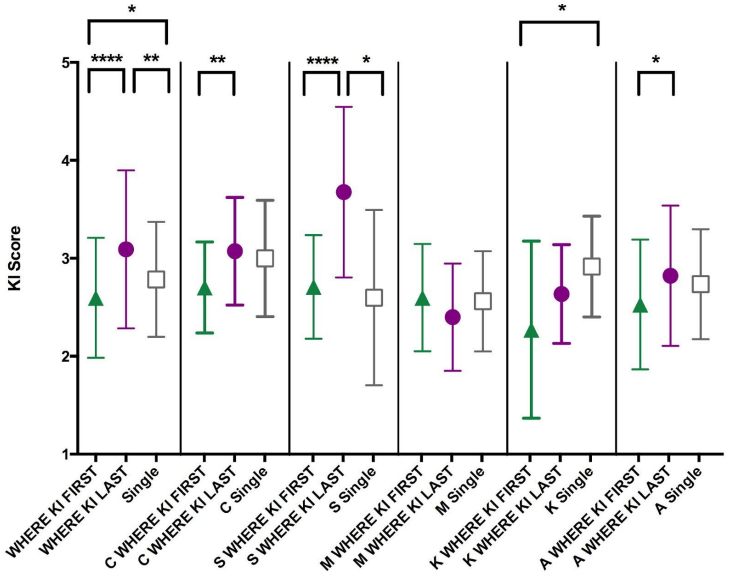


Figure 13. KI scores for students for the *Where* item across teachers

Table 7. Revision rate data for the KI portion of the *Where* item

Where KI	Submission	N	Average	St. Dev.	Cohen's d	Student's t-test (paired, two-tailed)	Revision Rate
All Teachers	First (revisions)		2.60	0.61			
	Last (revisions)	107	3.09	0.81	0.17	< 0.0001	60.45%
	Single	70	2.79	0.59			
C	First (revisions)		2.70	0.47			
	Last (revisions)	27	3.07	0.55	0.18	0.0021	60.00%
	Single	18	3.00	0.59			
S	First (revisions)		2.71	0.53			
	Last (revisions)	31	3.68	0.87	**0.35**	< 0.0001	86.11%
	Single	5	2.60	0.89			
M	First (revisions)		2.60	0.55			
	Last (revisions)	5	2.40	0.55	-0.09	0.3739	23.81%
	Single	16	2.56	0.51			
K	First (revisions)		2.27	0.90			
	Last (revisions)	11	2.64	0.50	0.13	0.1039	47.83%
	Single	12	2.92	0.51			
A	First (revisions)		2.53	0.66			
	Last (revisions)	34	2.82	0.72	0.11	0.0230	64.15%
	Single	19	2.74	0.56			

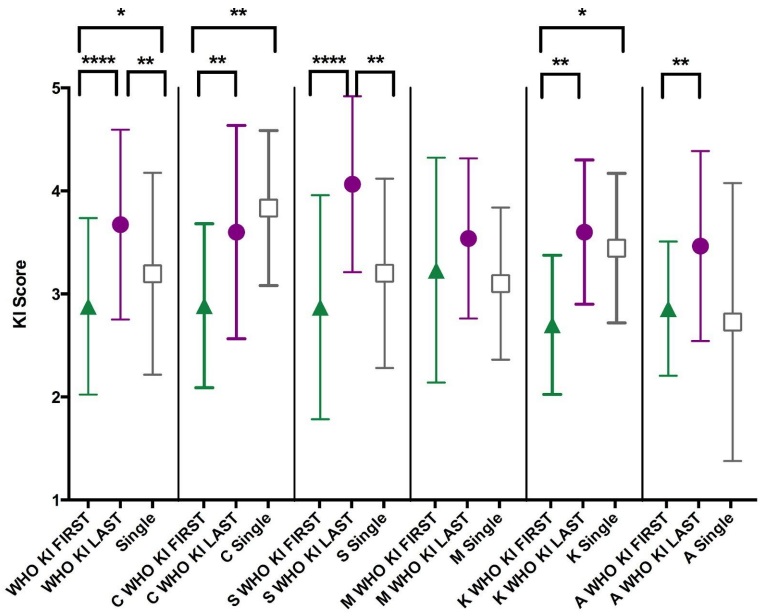


Figure 14. KI scores for students for the *Who* item across teachers

Table 8. Revision rate data for the KI portion of the *Who* item

Who KI	Submission	N	Average	St. Dev.	Cohen's d	Student's t-test (paired, two-tailed)	Revision Rate
All Teachers	First (revisions		2.88	0.86			
	Last (revisions	116	3.67	0.92	**0.22**	< 0.0001	71.60%
	Single	46	3.20	0.98			
C	First (revisions		2.89	0.80			
	Last (revisions	35	3.60	1.03	**0.20**	0.0010	85.37%
	Single	6	3.83	0.75			
S	First (revisions		2.87	1.09			
	Last (revisions	31	4.06	0.85	**0.31**	< 0.0001	75.61%
	Single	10	3.20	0.92			
M	First (revisions		3.23	1.09			
	Last (revisions	13	3.54	0.78	0.08	0.2643	56.52%
	Single	10	3.10	0.74			
K	First (revisions		2.70	0.67			
	Last (revisions	10	3.60	0.70	**0.33**	0.0100	52.63%
	Single	9	3.44	0.73			
A	First (revisions		2.86	0.65			
	Last (revisions	28	3.46	0.92	0.19	0.0032	71.79%
	Single	11	2.73	1.35			

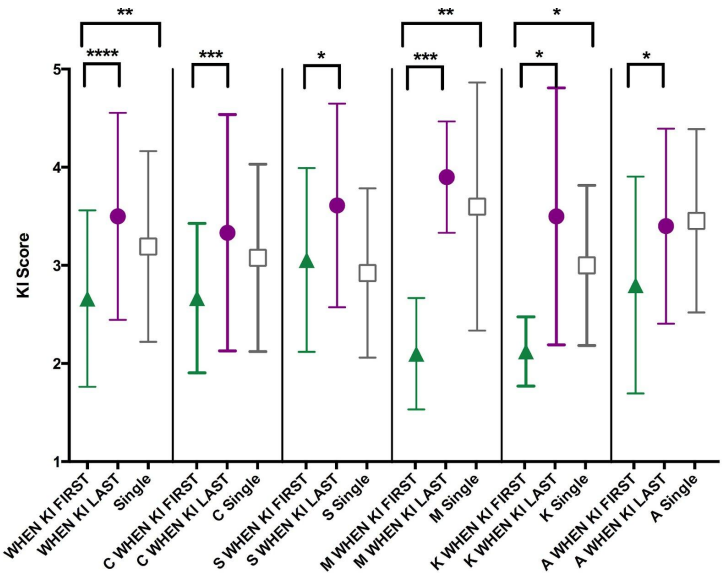


Figure 15. KI scores for students for the *When* item across teachers

Table 9. Revision rate data for the KI portion of the *When* item

When KI	Submission	N	Average	St. Dev.	Cohen's d	Student's t-test (paired, two-tailed)	Revision Rate
All Teachers	First (revisions		2.66	0.90			
	Last (revisions)	80	3.50	1.06	**0.21**	< 0.0001	58.39%
	Single	57	3.19	0.97			
C	First (revisions		2.67	0.76			
	Last (revisions)	24	3.33	1.20	0.17	0.0006	64.86%
	Single	13	3.08	0.95			
S	First (revisions		3.06	0.94			
	Last (revisions)	18	3.61	1.04	0.14	0.0204	58.06%
	Single	13	2.92	0.86			
M	First (revisions		2.10	0.57			
	Last (revisions)	10	3.90	0.57	**0.79**	0.0002	50.00%
	Single	10	3.60	1.26			
K	First (revisions		2.13	0.35			
	Last (revisions)	8	3.50	1.31	**0.41**	0.0203	44.44%
	Single	10	3.00	0.82			
A	First (revisions		2.80	1.11			
	Last (revisions)	20	3.40	0.99	0.14	0.0298	64.52%
	Single	11	3.45	0.93			

We are continuing our analysis of the revision data. Other questions that we would like to investigate include: Did students who revised do better on the post-test than those who did not? Did students in experimental classes revise more than those in control classes?

Discussion and Research Directions

Our LOOPS investigation presents good evidence for the benefit of a real-time formative feedback system. Both teachers and students benefited from use of the curriculum and technology.

We found that teachers, given some practice, were very capable of using the LOOPS system in real time. All teachers developed some agility in reacting to student work and formative feedback data, which included automated scoring of graph responses and summary information representing student level of effort and progress through activities. Making student work visible to teachers and enabling teachers to share that work with students greatly facilitated both of the teaching patterns in our study—holding class discussions and working with individuals. The formative feedback data enabled teachers

quickly and easily to determine when to hold class discussion and what to discuss, e.g., choosing a topic that showed variation in student responses. The data also enabled teachers quickly and easily to identify and work with students who needed help. We found that the LOOPS system accommodated a variety of teaching styles, with each teacher able to find a comfortable balance between working with individuals and discussing student work and curricular activities with the class as a whole.

The teachers needed minimal professional development in learning to use the technology. As mentioned, they could determine *when* to hold class discussion and *what* to discuss. What proved difficult for some, however, was *how* to conduct a class discussion based on student ideas and work. These teachers would benefit from professional development that focuses on teacher-student discourse and patterns of classroom talk (e.g., Shemwell & Furtak 2010; Ruiz-Primo & Furtak 2007; Penuel et al. 2012; Edwards & Westgate 1994).

We found that students benefited from the use of LOOPS with both teaching patterns. In learning science content and in developing skill at crafting evidence-based explanations, students showed significant gains on all pre-post test items across all classes and all teachers. We did not find measureable differences in student pre-post learning gains with one of the patterns over the other. We are continuing to analyze our embedded assessment data.

Both teaching patterns encouraged students to revise their work, creating a classroom culture in which students were comfortable discussing those revisions. We found via analysis of embedded assessments that students who revised their work improved their explanations and gave stronger answers than students who did not.

We see many possibilities for further research on LOOPS or LOOPS-like formative feedback systems. One next step is to repeat the LOOPS study with a larger N, using our current curriculum, developing a new curriculum, or adapting an existing curriculum. Adapting an existing curriculum has the advantage of a built-in control group of teachers and students already using the curriculum.

We also would like to expand the LOOPS notion of teaching pattern to include finer-grained classroom patterns that take into account specific characteristics of classroom discourse. Such patterns include, for example, those articulated by Linn & Eylon (2011); Penuel et al. (2012); Ruiz-Primo & Furtak (2007); Shemwell & Furtak (2010); and Van de Pol, Volman, & Beishuizen (2011).

Another research direction that we advocate is that of extending LOOPS to work with "freer" forms of curricula, such as project- or problem-based curricula, that are not organized as a set of linear activities. Such an extension would enable a teacher to use formative feedback to easily modify instruction and activities "on the fly" or differentiate instruction to better meet the needs of individual students. Investigations into new sorts of summary information that could give teachers insight into their students' understanding would complement this research.

Conclusion

Technology-enabled formative feedback systems such as LOOPS hold great promise in improving teaching and learning, especially in science where "…instructional responsiveness is a crucial aspect of scientific inquiry teaching" (Ruiz-Primo & Furtak 2007, p. 78). Instructional responsiveness is afforded by giving teachers tools to gain insight into their students' thinking and to engage their students in reflective conversation. LOOPS provides examples of such tools—the wireless collection of student work, an interface that enables teachers to view student work and summary assessment data and choose student work for public display, and a means for gauging both class and individual progress. Its delivery

on a mobile computer increases its effectiveness, as teachers can access formative feedback data while circulating among their students, identifying and working with individual students who may need help. Its integration with a curriculum that includes computational models and instruments such as motion probes ensures that students are engaged in activities that will help develop inquiry skills, and that assessment data for both process and product can be investigated. LOOPS is an example of a formative feedback system that can be used effectively to guide instruction and improve student learning in science. We look forward to continuing with its research and development.

References

Bell, B. & Crowe, B. (2001) Formative assessment and science education. Dorbrecht: Kluwer.

Black, P. (1993) Formative an summative assessment by teachers. Studies In Science Education, 21, 49-97.

Black, P. & William, D. (1998) Assessment and classroom learning. Assessment in Education, 5, 7-74.

Duschl, R.A. Assessment in inquiry. In J.M. Atkin & J.E. Coffey (eds) Everyday assessment in the science classroom, 41-59. Washington, DC: National Science Teachers Association Press.

Edwards, A.D. & Westgate, D.P.G. (1994) Investigating classroom talk (2nd ed.). London: Westgate.

Furtak, E.M., Ruiz-Primo, M.A., Shemwell, J.T., Ayala, C.C., Brandon, P.R., Shavelson, R.J., & Yin, Yue. (2008) On the fidelity of implementing embedded formative assessments and its relation to student learning. Applied Measurement in Education, 21 (4), 360-389.

Gerard, L.F., Spitulnik, M., & Linn, M.C. (2010) Teacher use of evidence to customize inquiry science instruction. Journal of Research in Science Teaching, 47 (9), 1037-1063.

Linn, M. C., & Eylon, B.-S. (2011) Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration. New York: Routledge.

Penuel, W. R., Moorthy, S., DeBerger, A., Beauvineau, Y., & Allison, K. (2012). Tools for Orchestrating Productive Talk in Science Classrooms. The Future of Learning: Proceedings of the 10th International Conference of the Learning Sciences (ICLS 2012). Sydney, Australia: International Society of the Learning Sciences.

Ruiz-Primo, M. A., & Furtak, E. M. (2007) Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. Journal of Research in Science Teaching, 44 (1), 57-84.

Shemwell, J.T. & Furtak, E.M. (2010) Science classroom discussion as scientific argumentation: a study of conceptually rich (and poor) student talk. Educational Assessment, 15 (3), 222-250.

Slotta, J.D. & Linn, M.C. (2000) The knowledge integration environment: helping students use the internet effectively. In P.A. Alexander & P.H. Winne (eds) Handbook of Educational Psychology, 511-544. Mahwah, NJ: Lawrence Erlbaum Associates.

Van de Pol, J., Volman, M., & Beishuizen, J. (2011) Patterns of contingent teaching in teacher-student interaction. Learning and Instruction, 21, 46-57.

Appendix

Alternate view of data in Table 3 for pre-post test data for five of the 13 items for control and experimental groups is shown in Figure 1.

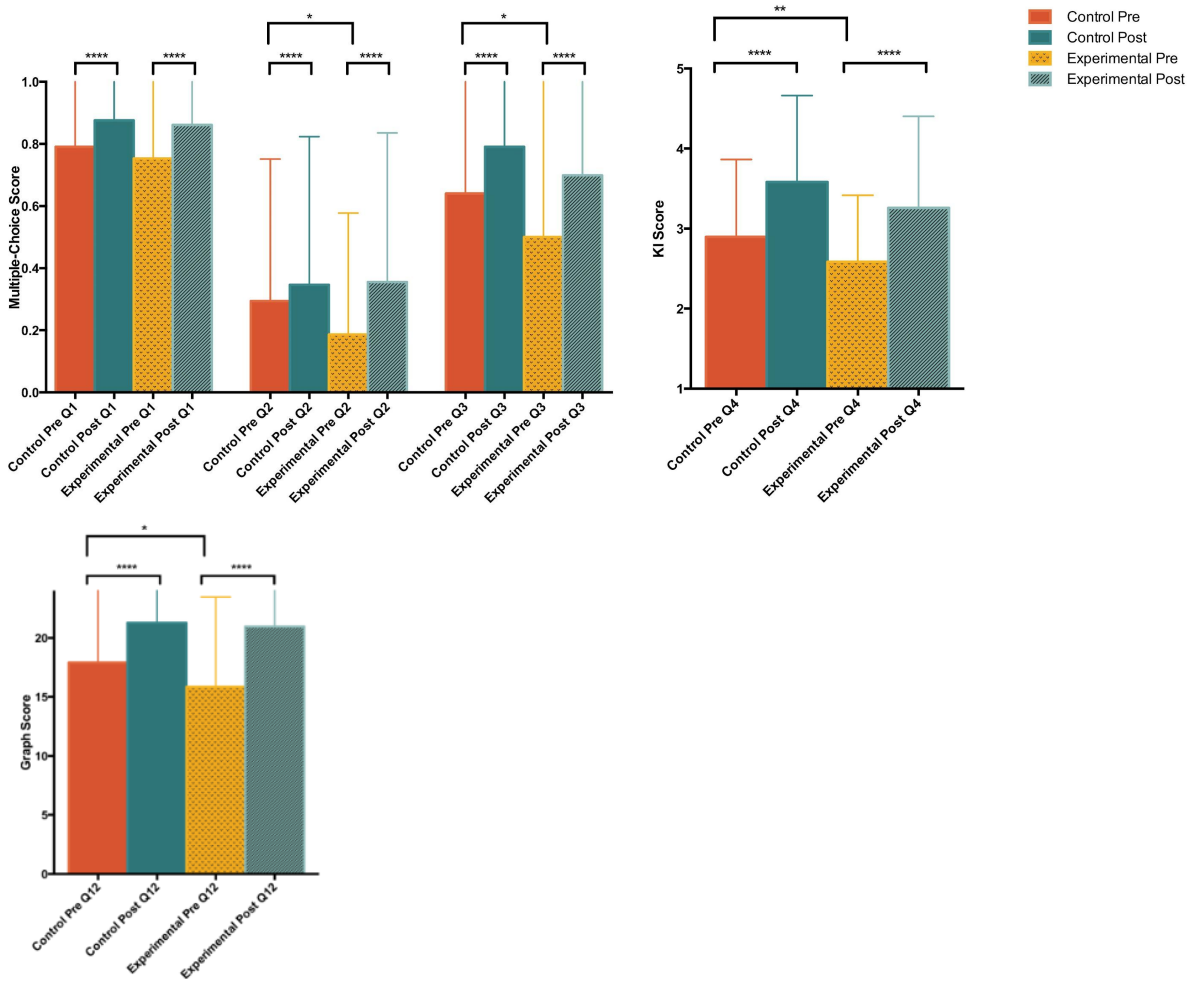


Figure 1. Pre-post data for control vs experimental classes for items 1 to 3 (multiple choice), 4 (open response), and 12 (graphing); $p<0.0001$ (****); $p<0.001$ (***); $p<0.01$ (**); $p<0.05$ (*)

An alternate view of the pre-post data shown in Figure 9 for all teachers for graphing item 12 is shown in Table 1.

Table 1. Pre-post data for all teachers, control vs. experimental groups, for the graphing item (12): data show learning gains across all teachers' classes, both control and experimental groups.

	All Control					All Experimental					All Teachers: Control Pre vs. Experimental Pre
	N	Average	Std. Dev.	Paired, two-tailed Student's t-test	Cohen's d (SD)	N	Average	Std. Dev.	Paired, two-tailed Student's t-test	Cohen's d (SD)	
12 Pre	136	17.91	6.99			158	15.84	7.63			
12 Post	136	21.29	5.34	< 0.0001	1.362	158	20.97	5.32	< 0.0001	2.015	0.0165
	C Control					C Experimental					C: Control Pre vs. Experimental Pre
	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	
12 Pre	45	18.93	5.64			41	15.32	7.10			
12 Post	45	22.17	3.43	< 0.0001	1.521	41	21.15	4.93	< 0.0001	2.377	0.0151
	S Control					S Experimental					S: Control Pre vs. Experimental Pre
	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	
12 Pre	37	17.59	7.12			36	16.33	8.19			
12 Post	37	21.76	4.94	< 0.0001	1.695	36	21.69	4.68	< 0.0001	2.113	0.4845
	M Control					M Experimental					M: Control Pre vs. Experimental Pre
	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	
12 Pre	18	20.17	5.87			21	17.38	7.83			
12 Post	18	20.89	6.76	0.6679	0.287	21	21.57	5.57	0.0128	1.619	0.223
	K Control					K Experimental					K: Control Pre vs. Experimental Pre
	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	
12 Pre	14	20.93	5.23			16	19.38	5.49			
12 Post	14	22.86	3.21	0.1078	0.939	16	21.38	5.43	0.0484	0.856	0.4356
	A Control					A Experimental					A: Control Pre vs. Experimental Pre
	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	N	Average	Std. Dev.	Paired, two-tailed	Cohen's d (SD)	
12 Pre	22	12.59	8.53			44	13.91	7.88			
12 Post	22	18.00	7.67	0.0027	1.900	44	19.03	6.61	< 0.0001	1.903	0.5355