Eyes on the Prize: Considering How Design Research Can Lead to Sustainable Innovation

Vandana Thadani\textsuperscript{1}, Melissa Cook\textsuperscript{2}, Kelli A. Millwood\textsuperscript{2}, Aletha Harven\textsuperscript{2}, Deborah Fields\textsuperscript{2}, Kathy Griffis\textsuperscript{3}, Joe Wise\textsuperscript{4}, Karen Kim\textsuperscript{5}, William A. Sandoval\textsuperscript{2}

\textsuperscript{1}Department of Psychology, Loyola Marymount University, Los Angeles, CA.
\textsuperscript{2}Graduate School of Education & Information Studies, University of California, Los Angeles.
\textsuperscript{3}Science Department, The Buckley School, Sherman Oaks, CA.
\textsuperscript{4}Center for Effective Learning, New Roads School, Santa Monica, CA.
\textsuperscript{5}Center for Embedded Networked Sensing, University of California, Los Angeles.

Correspondence to:
William A. Sandoval
2339 Moore Hall, Box 951521
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1521
Phone: (310) 794-5431
FAX: (310) 206-6293
sandoval@gseis.ucla.edu
Abstract

Research on computer-supported environments for science learning has been going on successfully for nearly three decades. Yet, most learning environments research involves “hothouse” projects where researchers and developers are intimately involved in implementation. The success of such environments is well documented, but their reach is severely limited in schools. This symposium presents empirical findings from the first year of such a research project that aims to increase its reach to national scale. We ask what kinds of data are needed to properly understand local adaptations of designed learning environments, and how can research in the early phases of development contribute to educative, sustainable curricular materials. Our findings and the challenges they raise have implications for learning environment design research.
Introduction

What will it take to make rich, provocative learning environments available to all students, especially students in urban schools? Over the last two decades, a number of learning environments have been developed and successfully implemented in various sorts of classrooms, especially for science (Edelson, Gordin, & Pea, 1999; Hickey, Kindfield, Horwitz, & Christie, 2003; Linn, Bell, & Hsi, 1998; White, 1993). These research-based efforts, however, have almost exclusively been “hothouse” projects, where a research team works closely with one or a small number of teachers to develop and refine an intervention through cycles of design, implementation, analysis, and revision. Such projects prove the concept that well-designed learning environments can not only help students learn but profoundly change the nature of learning in classrooms. Yet, it remains problematic for such learning environments to succeed once researchers have left the classroom, despite the long recognition of this problem (e.g., Collins, 1992). This lack of knowledge about how to make research-based learning environments sustainable, much less scalable, presents a serious challenge to learning environment researchers.

Within a policy context of standards-based accountability, hothouse projects become less and less tenable. They often do not measure outcomes in ways that policymakers and the public can interpret in relation to standardized tests. In science, these projects often aim for ambitious goals, such as helping students to learn scientific inquiry, that have proven difficult to assess in statistically adequate ways. Moreover, research-based learning environments are so expensive and time-consuming to build, that it is imperative that the research community understand how variations in their use can be accommodated, and adaptations guided towards stable, productive implementations. At the same time, there is a constant demand for innovation – for increasing the range of opportunities that students and teachers have to inquire into the natural world. The
question we address in this paper is how to frame the research and development trajectory of a learning environments project that can help to move it effectively from the hothouse to widespread use.

**Learning science through inquiry**

Inquiry has been advocated as the proper core of good science instruction for decades (National Research Council, 1996; Schwab, 1962). As important as inquiry has become in science education, it remains difficult for most students and may be more challenging for some student groups over others. Studies of inquiry learning have demonstrated that students have trouble devising scientific questions, designing experiments, interpreting data, drawing conclusions, and constructing scientific arguments based on their data (see the brief review in Sandoval, 2005). The challenges that students experience in inquiry classrooms may differ based on their racial, ethno-linguistic, and socio-economic backgrounds (Lee, 2003; Lee, Deaktor, Hart, Cuevas, & Enders, 2005; Songer, Lee, & McDonald, 2003). For instance, ethno-linguistic minorities may have a particularly difficult time mastering scientific discourse (such as argumentation) because their cultural backgrounds and norms expect very different behaviors from the norms of scientific practice, a gap which might make learning scientific discourses particularly inaccessible for these students. At the same time, Songer et al. (2003) have argued that minority students, in particular, can benefit from inquiry-based instruction because it challenges the teacher-directed, fact-based “pedagogy of poverty” (p. 131) that commonly exists in classrooms that serve them.

Research on inquiry-oriented science interventions has typically focused on particular kinds of learning outcomes, especially science content learning and learning of inquiry skills, such as designing experiments. We know of no comprehensive reviews of this research that summarizes learning outcomes, although there are certainly a number of exemplars of successful
interventions. The National Research Council recently conducted a partial review of this work and found, especially in contrast to typical laboratories, that inquiry-oriented instruction is more effective at improving students' understanding of targeted science concepts and certain targeted inquiry skills (National Research Council, 2005). This report also found, however, that a number of other outcomes of interest have been neglected in inquiry research. For example, researchers' generally have not examined whether or not inquiry experiences in school improve students' attitudes toward or interest in science. Another important outcome is students' understanding of the nature of science. Here, the evidence is that inquiry experiences in themselves are insufficient to change students' ideas about the nature of scientific knowledge and how that knowledge is produced (Sandoval, 2005).

**Teaching science through inquiry**

Inquiry teaching is rare, and seems to have always been so, despite recurring calls for inquiry teaching over the last fifty years (Schwab, 1962; Tobin, Tippins, & Gallard, 1994; Welch, Klopfer, Aikenhead, & Robinson, 1981). A crucial reason for this is that few teachers have personal experience with scientific inquiry, hence little idea of how to enact inquiry with students (Windschitl, 2003). Moreover, teacher preparation does not appear to equip most science teachers with much knowledge about common student ideas about science topics, despite a huge amount of research on student conceptions. Science teachers thus face a major learning challenge from inquiry-based reforms that includes understanding student conceptions, understanding the nature of scientific inquiry, and learning to develop students’ ideas through inquiry instruction.

Curriculum-based interventions have the potential for helping teachers to deliver rich, inquiry-based science instruction to diverse groups of students. Well-designed curricular materials can be a mechanism for sharing the “practitioner knowledge” (Hiebert, Gallimore,
Stigler, 2002, p. 4) that teachers must use for teaching – particularly when they attempt to depart from more teacher-directed, lecture-based instruction. Such curricula can provide this knowledge without each teacher having to “reinvent” their own instructional strategies from scratch (Ball & Cohen, 1996; Davis & Krajcik, 2005; Hiebert et al., 2002). It may very well be the case that educative curricula are insufficient to support the teacher learning needed to transform science instruction towards inquiry. The demands on teacher preparation for such reforms seem great (see National Research Council, 2005). Still, the need for educative curricula remains, as does the need to understand more thoroughly what makes curricula educative for a broad array of teachers.

Given the current state of research on inquiry-oriented science environments, we have tried to frame as many of these issues as possible at the start of our research and development trajectory. We are developing guided inquiry units for use in middle or high school life science courses. There are several features our effort shares with others, and some features that are unique. We describe our current work below to provide context for the study reported here. For the moment, we note that our research approach at our current early stage of development has been to concurrently collect data on a range of student outcomes including content learning, attitude toward science, interest in inquiry specifically, practices of argumentation, and conception of science. At the same time, we are studying variations in teacher implementation with the dual aims of linking such variations to student outcomes and to developing educative materials for teachers.

**Aims of design-based research**

Our approach is grounded in what has come to be known as design-based research. Design-based research simultaneously pursues dual aims of a) formative evaluation of learning environments to improve their design; and b) research on complex learning in naturalistic
settings (Design-Based Research Collective, 2003). In our case, we are interested in both student and teacher learning. What do students learn through their experiences with our curricular units? What do teachers learn about science and science teaching as they learn to teach these units? One of the difficulties of design-based research is that the learning environments being designed are usually creating the conditions for the learning that you want to study. For instance, we cannot simply ask how do students learn to do scientific inquiry. We have to first create the conditions that allow them to do scientific inquiry. Consequently, figuring out how they learn to do inquiry is conflated with the conditions created for doing so. We would argue that this is generally true of intervention research of any sort; design-based research merely makes that conflation explicit. The burden this places on design-based researchers is to make explicit their conjectures about how specific designs will support specific kinds of learning. Furthermore, such conjectures have to be projected forward through a trajectory of research and development.

Our own experience in such efforts is that how to map theoretical conjectures about learning, whether we are talking about students or teachers, through a trajectory of design and research cycles is poorly understood. We make no claims about knowing more about how to do this work than others. On the contrary, one goal of this study is to uncover particular challenges to this kind of work. At the same time, we believe our effort to be more comprehensive in our research at a very early stage of development is new, and has the potential to improve the long-term results.

**Sensing the Environment**

The Center for Embedded Networked Sensing (CENS), a Science and Technology Center funded by NSF, conducts research to develop and deploy sensor arrays to collect real-time data of natural phenomena for use in scientific exploration. To help students and teachers engage with
such data in meaningful ways, we are designing curricular materials aligned to both national California science standards. To align with California standards in a way that enables to provide a series of inquiry experiences throughout a single school year, we have focused on 7th grade life science topics. Our approach is to develop open-ended modules of 3 to 4 weeks' duration that cohesively address a number of interrelated topics and standards.

The *Sensing the Environment* module calls for students to investigate the relationship between environmental variations and plant adaptations using temperature, humidity, and light intensity data from a sensor network embedded in a local mountain range. Students must reason about the relationship between plant structure and function to construct a data-supported explanation of why plants look different in different regions of the mountains. The module includes several activities that guide students’ understanding of photosynthesis, transpiration, and evolution, including: 1) staging activities that develop the concepts and skills necessary for students to conduct their own inquiry, 2) guided inquiry activities that scaffold students through exploration and experimentation using web-delivered data from sensors in the field, and 3) synthesis activities that engage students in the integration of their own investigations with subject-matter content and the nature of scientific knowledge and practice.

Curricular reforms are more likely to succeed if they simultaneously provide teachers with well-designed curricula, knowledge that can help teachers anticipate student thinking, and instructional strategies (e.g., guiding questions, strategies for helping students collect and analyze data or for assisting students to organize notes) for promoting inquiry (Davis & Krajcik, 2005). Recognizing that teaching with inquiry is difficult for most teachers, and that interventions frequently fall short because of inadequate training and support (Desimone, 2002), this unit was built with several mechanisms for supporting teachers’ efforts to use it. Teachers
were provided a detailed teacher’s guide that helped them anticipate what to expect during the unit. Included in the guide were questions that teachers could use to lead students through an exploration of the content, common student responses to questions, common student conceptions, strategies for reinforcing key concepts and identifying variables, and approximate timelines for activities. Much of this information had been compiled through extensive and iterative piloting of curriculum materials. We also provided a one-day face-to-face training workshop during which one of the curriculum developers modeled teaching with the unit while teachers participating in the study performed investigations as if they were students. Teachers watched and critiqued video of pilot teachers teaching the unit, and discussed their concerns and strategies for classroom inquiry.

The Sensing the Environment unit consisted of eight major activities, design to span approximately 15 days of instruction (assuming 50 minute lessons). The unit asked, "Why do plants look different?" On the first day, students looked at a picture of the area where the sensor network they would use later was placed. They noticed that the plants in this general area varied, as we intended, and that these variations were visible because of differences in leaf structure (size, shape, and texture). This was followed by a sequence of activities designed to help students learn the functions of leaves, especially photosynthesis and transpiration, as they would need this knowledge to carry out their online investigations. After exploring sensor data in small groups via the web, students individually wrote essays and then reviewed each others' work. The last day of the unit summarized what students had learned and produced a consensus answer to the driving question.
Methods

We were interested in a number of questions about the efficacy of *Sensing the Environment*. Would students learn more about targeted science concepts through our unit than from typical instruction? Would students' participation in our unit improve their general attitudes toward science compared to typical instruction? Would students' participation influence their interest in and understanding of scientific inquiry? Besides these questions we also wanted to explore these students' ideas about scientific argumentation and how such ideas influenced their work. Finally, we wanted to document potential variations in teachers' implementations of the unit and, ultimately, link differences in implementations to student outcomes with the aim of improving our materials. Addressing these questions demanded a range of methods. We approached the efficacy questions through a quasi-experimental design comparing paired classes at a number of schools either using our unit or typical instruction on the same topics. Other questions were approached using sub-samples from the larger sample, as described individually below.

Overall context and participants

This study represented the first field test of *Sensing the Environment*. We wanted to see the unit implemented in a range of schools serving diverse student populations while keeping the number of teachers relatively small. We recruited teachers through our own personal contacts, through the CENS education advisory board, and through professional development programs at UCLA and other local institutions. We recruited ten teachers from five middle schools throughout urban Los Angeles. Basic demographics of each school are shown in Table 1 (Los Angeles is a highly diverse city and there are small, variable ethnic populations at each school that do not fall into the common ethnicities reported in the table, thus those percentages do not sum to 100).
Table 1. Demographics of participating schools. All numbers are percentages.

<table>
<thead>
<tr>
<th>School</th>
<th>African-American</th>
<th>Asian-American</th>
<th>Caucasian</th>
<th>Latino</th>
<th>Free or Reduced Lunch</th>
</tr>
</thead>
<tbody>
<tr>
<td>School1</td>
<td>25</td>
<td>8</td>
<td>9</td>
<td>57</td>
<td>75</td>
</tr>
<tr>
<td>School2</td>
<td>65</td>
<td>4</td>
<td>23</td>
<td>7</td>
<td>36</td>
</tr>
<tr>
<td>School3</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>71</td>
<td>3</td>
</tr>
<tr>
<td>School4</td>
<td>1</td>
<td>10</td>
<td>74</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>School5</td>
<td>33</td>
<td>–</td>
<td>–</td>
<td>65</td>
<td>43</td>
</tr>
</tbody>
</table>

*a* includes Asian, Filipino, and Pacific Islander

Two teachers from School1, four from School2, two from School3, and one each from School4 and School5 participated in this study. All received a stipend for their participation. We attempted to recruit teachers in pairs for our quasi-experimental design. Since we were unable to do this at Schools 4 and 5, these were not included in the quasi-experimental portions of the study.

**Science learning**

Teachers were recruited from each school in pairs, with the understanding that if they volunteered for the study, they would be randomly assigned to either the CENS or a comparison condition. School was a stratum for random assignment: for every pair of teachers who participated from each school, one was assigned randomly to the CENS condition and the other served as his/her control. Furthermore, each teacher selected one “target” class period for follow-up in the study. We successfully recruited pairs of teachers at Schools 1, 2, and 3.

The CENS learning assessment is a closed-ended, paper-and-pencil test that we developed, that is aligned to each CA content standard targeted by this study. It assesses students’ content knowledge of photosynthesis, transpiration, and evolution as well as their knowledge about the process of scientific investigation (known in the California state standards, and referred to below, as Investigation and Experimentation [IE]). We created this learning assessment primarily because we needed assessment items that were closely linked to the CA content standards targeted by the study. To ensure that the assessment did not unfairly favor
CENS students, it was reviewed by a panel of four science education experts -- university-level science-education faculty who had experience in science-education reform efforts, as well as several middle school science teachers. Their feedback, coupled with additional feedback gathered during a pilot of all materials, was used to revise the instrument for use in this study. In its final version, 14 points represented content items (photosynthesis, transpiration, and evolution) and 10 points represented scientific investigation items, for a total of 24 points on the assessment as a whole. This assessment was used as a pre- and post-test of student knowledge on the relevant content standards.

This test was given before and after the unit to students in all five of the schools where the CENS unit was implemented. For purposes of judging the efficacy of our unit in comparison to teachers' typical instruction on these topics, we limit our presentation of results below to the three schools where we had comparison classes.

**Attitudes toward science**

We developed a 12-item self-report questionnaire on attitudes toward science by adapting several items from Eccles and Wigfield's (1995) questionnaire of students’ attitude toward mathematics, and adding others. Items tapped two of the dimensions of attitude towards science identified by Eccles and Wigfield: perceived value of science (e.g., “I can use the things I learn in science out in the real world.”) and intrinsic interest in science (e.g., “Science is boring.”), with 6 items addressing each dimension. Items on each scale were averaged to generate a single Interest and Perceived value score for each student, with low ratings indicating low perceived value/interest and high ratings suggesting high perceived value/interest. All students in the study were given this questionnaire. As with the learning assessment, a primary focus of our analyses presented below is on the comparison between CENS students and their comparable peers.
This general attitude survey, however, is ill-suited to elicit students' perceptions of the value of inquiry specifically. We therefore developed a novel instrument to try to find out how students valued the kinds of inquiry tasks that they were asked to perform during the unit. Students were asked to rate four tasks that were central to the unit: (1) deciding which data they needed and how to collect it, (2) gathering and analyzing data, (3) using data to support their claims, and (4) writing an explanation (essay) about their data. Five-point Likert scales were used to rate students responses to the following questions: (a) was this different from what you normally do in your science class?, (1-not very different to 5- very different), (b) compared to what you normally do in your science class, how interesting was it to decide which data you needed and how to get it?, (1-not very interesting to 5-very interesting), (c) is it useful to decide which data you need and how to get it?, (1-not very useful to 5-very useful), and (d) Compared to the work you normally do, how hard is it to decide which data you needed and how to get it?, (1- not very hard to 5-very hard). Additional space was provided under each question for students to explain their ratings. This survey was given to students at the end of the module. We report the number of students who responded to this survey in the results section to follow.

**Students' ideas about argumentation**

Sandoval (2005) has recently argued for the need to study the "practical" epistemological ideas that students have about science. His hypothesis is that students' ideas about their own knowledge-making practices may be quite different than their ideas about scientists' knowledge-making practices. We took this field test as the opportunity to begin to explore this hypothesis. As part of their work in the CENS unit, students individually wrote essays to answer the unit's driving questions of why plants look different. In this essay, students were expected to use the data they generated through their collaborative investigations. We collected all of the individual essays written by students in School5. This yielded essays from 33 students (20 boys, 13 girls).
We scored the quality of the arguments students made in their essays following a protocol developed in previous work (Sandoval & Millwood, 2005). We particularly wanted to know if students would write explanations in the conceptual terms targeted by the unit: 1) describing the functions of leaves, 2) how those functions were related to leaf structure, 3) how environmental variations could affect those structures, and 4) the differential fitness of structural variations in relation to environmental variables.

Student essays were scored in terms of these four components as follows. Students received one point for each of the four components they articulated in their essays, up to a maximum of 4. Then, for each component the degree of warrant that students provided was scored. Warrant refers to the level of justification students provided for their claims. For example, to warrant a claim that leaf size varies according to differences in temperature, students would have to cite temperature data and leaf size from at least two different sensor locations. Function claims did not require a warrant because the primary functions of a leaf, photosynthesis and transpiration, were considered shared knowledge in the classroom, and none of the data available in the investigation environment could be used to support claims of function. For other articulated claims, students' level of warrant was scored on a 4 point scale where 0 reflected no warrant at all, and 3 reflected a full warrant. Differences between levels will be explained in the results section. Higher scores on essays, especially on warrants, reflects more scientific forms of argumentation.

We also interviewed each of the 33 students about their essays following the unit. The interview began by asking students why they were writing their essay, what they thought their teacher’s goals were for the essay, and what their own goals were. The purpose of these questions in the first part of the interview was to ground the interview in the student’s work.
Next, each student was given a highlighter and asked to mark off all of the claims they made in their essay. All of the students were told that a claim is a statement that talks about what causes something. After the student marked off her claims she was asked which claim she was the most certain about, and why she was certain of that claim. Next the student was asked which claim she was the most uncertain about and why. Finally, each student was asked to describe the best way to convince someone of something in science, and why. The intent of these questions was to assess students' ideas about how scientific claims are best warranted. The questions about their own claims intended to elicit their criteria for justifying their own claims. The question about generally persuading someone "in science" was intended to compare their criteria for their own (school) work to their ideas about appropriate criteria in science. This interview provided us with data both about how students saw their own work, and how their ideas about proper warrants related to their actual practices of justifying claims in their essays.

Teacher adaptation

We collected two forms of data to document ways in which teachers implemented the CENS unit. Structured classroom observations were used to rate the extent to which classrooms could be characterized as practicing inquiry-based teaching and learning. First, a checklist of 11 items described events -- such as “Students generated predictions/hypotheses” or “Students explicitly applied learning from class to outside world” -- indicating that students were engaged in active learning, scientific thinking, and scientific processes. Several of these events should be characteristic of classrooms in which students were engaged in inquiry. Items on the checklist were simply marked yes or no whether or not they occurred at all during an observed lesson. Observers rated 6 additional indicators of active engagement in scientific thinking and processes (e.g., “The students made statements or asked questions that suggested they were thinking metacognitively.”). These six items were rated on two scales: Classrooms were rated on the
frequency of such events during an observed lesson (1=not at all characteristic; to 5= highly characteristic), and on the proportion of students involved in these events (1 = none to 5 = 75% or more). A third set of items described events characteristic of a teacher-directed, primarily lecture-based lesson (e.g., “The students’ primary task was to memorize facts/procedures for later reproduction.”). These items were rated on the frequency scale described above. The same panel of science education experts who had reviewed our learning assessment also reviewed items on the observation protocol. To compute reliability, two observers completed the classroom observation protocol in four classrooms. Cohen’s Kappa for the Yes-No items was .67; which conventionally is regarded as “good” (Bakeman & Gottman, 1986). Pearson’s r was used to compute reliability on the 19 items rated on the 5-point scale, and was determined to be acceptable, r(74) = .73.

These structured observations were conducted in both CENS and Comparison classrooms. We use them to produce a coarse picture of the differences between classes. They are suggestive of whether or not anything like inquiry might be taking place in these classrooms, and are useful measure of whether or not the instruction in CENS classrooms was different than in comparison classrooms. They are insufficient, however, to provide the data needed to understand the specific challenges that teachers face in implementing a novel, innovative unit of instruction.

To get a better picture of how teachers varied in their implementations, we videotaped three classrooms as they implemented the CENS unit. We recorded almost every day of instruction for each classroom. Two researchers reviewed each tape, creating an activity log to outline the major shifts in activity and participation structure in each lesson. We then selected two classrooms to compare for this preliminary study based on our impression that the two
classrooms differed in some ways in terms of how students and their work were framed, and based on the difference in learning gains for the two classrooms. Following the method of microanalysis of interaction (Erickson, 1992), we examined the activity logs for signs of significant events in each lesson. We then transcribed these events. We are at a nascent stage in this analysis, but will present our initial findings of interest below.

**Results**

We present our results in several parts, corresponding to the various questions we aimed to answer in this study. Various measures were administered to students across several days. Absenteeism rates varied between the schools, but was quite common in most of them. Consequently, there is some variability in the samples for each measure. We describe sample sizes within each section.

**Science learning**

Across the three schools making up the quasi-experimental sample for this study, 150 students completed both the pre- and post- science test. Table 2 shows the distribution of students within schools and conditions. Data were analyzed separately at each school. Students' scores on the science learning assessment were analyzed using repeated measures ANOVA, with Time (pre v. post) as the within subjects factor, and Group (CENS v. Comparison) as the between subject factor. Partial-eta-squared (partial $\eta^2$) are reported as effect size estimates for all significant effects. According to Cohen (1988), effect sizes of .01 are conventionally considered small, .06 are considered medium, and .14 are considered large (though Cohen cautions against a rigid classification of effect size).
Table 2. Number of students compared for science learning in this study.

<table>
<thead>
<tr>
<th>School 1</th>
<th>School 2</th>
<th>School 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>14</td>
<td>32</td>
</tr>
<tr>
<td>CENS</td>
<td>11</td>
<td>41</td>
</tr>
<tr>
<td>TOTAL</td>
<td>25</td>
<td>73</td>
</tr>
</tbody>
</table>

There was a significant effect of Time on Content Scores at all three schools, $F(1, 23) = 8.21, p = .009$, partial $\eta^2 = .26$; $F(1, 71) = 41.26, p < .001$, partial $\eta^2 = .37$; $F(1, 50) = 19.69, p < .001$, partial $\eta^2 = .28$ (at Schools 1, 2, and 3, respectively). Table 3 displays means and standard deviations at pre- and post-test (collapsed across conditions) at all schools. As would be expected, students’ post-test scores were higher than their pre-test scores, though baseline scores at School 3 were quite a bit higher than those at Schools 1 and 2. At School 1, the interaction between Time and Group was not significant, $F(1, 23) = 2.74, p = .11$ In other words, the gain made by CENS students did not differ significantly from that made by Comparison students, though as Figure 1 illustrates, the relative gains made by the CENS group at this school were larger than those made in the other Schools. We observed a significant interaction between Time and Group at School 2, $F(1, 71) = 20.87, p < .001$, partial $\eta^2 = .23$, with CENS students making greater gains from pre- to post-test than their peers in the Comparison groups. At School 3, the interaction between Time and Group was not significant, $F(1, 50) = .05, p = .82$ (Figure 1).

Table 3. Descriptive statistics for learning assessment.

<table>
<thead>
<tr>
<th>School 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest Mean</td>
<td>Pretest SD</td>
<td>Posttest Mean</td>
<td>Posttest SD</td>
<td>Average Gain</td>
</tr>
<tr>
<td>Content Items</td>
<td>6.28</td>
<td>1.75</td>
<td>7.54</td>
<td>2.09</td>
<td>1.26</td>
</tr>
<tr>
<td>IE Items</td>
<td>4.96</td>
<td>2.04</td>
<td>5.72</td>
<td>2.00</td>
<td>.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>School 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest Mean</td>
<td>Pretest SD</td>
<td>Posttest Mean</td>
<td>Posttest SD</td>
<td>Average Gain</td>
</tr>
<tr>
<td>Content Items</td>
<td>7.68</td>
<td>2.11</td>
<td>9.27</td>
<td>2.87</td>
<td>1.59</td>
</tr>
<tr>
<td>IE Items</td>
<td>7.60</td>
<td>2.05</td>
<td>7.96</td>
<td>1.75</td>
<td>.36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>School 3</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest Mean</td>
<td>Pretest SD</td>
<td>Posttest Mean</td>
<td>Posttest SD</td>
<td>Average Gain</td>
</tr>
<tr>
<td>Content Items</td>
<td>9.71</td>
<td>2.08</td>
<td>11.00</td>
<td>2.03</td>
<td>1.29</td>
</tr>
<tr>
<td>IE Items</td>
<td>8.39</td>
<td>1.40</td>
<td>8.74</td>
<td>1.32</td>
<td>.35</td>
</tr>
</tbody>
</table>
Findings on IE items were not as robust, in part it appears, because many students were at ceiling on this part of the learning assessment at pre-test. At School 1, there was a trend for the effect of Time on IE scores, $F(1, 23) = 3.63, p = .07$, partial $\eta^2 = .14$. There was a significant effect of Time on IE scores at both School 2, $F(1, 71) = 4.95, p = .03$, partial $\eta^2 = .07$, and School 3, $F(1, 50) = 5.66, p = .02$, partial $\eta^2 = .10$. Again, post-test scores were higher than pre-test scores at all three schools (See Table 3). At Schools 1 and 3, there was no interaction between Time and Group, $F(1, 23) = .18, p = .68; F(1, 50) = .91, p = .35$, for Schools 1 and 3, respectively. At School 3, however, students were close to ceiling on the pre-test, making this lack of effect difficult to interpret. Finally, at School 2, there was a trend for an interaction between Time and Group, $F(1, 71) = 3.55, p = .06$, partial $\eta^2 = .05$. Comparison students’ scores increased more relative to those of CENS students; again, however, the difference in gain is difficult to interpret because CENS students were close to ceiling on the pre-test (Figure 2).

Figure 1. Mean Content Score (± SE) as a function of Time and Group.
Attitudes toward science

General attitudes about science

Between the three schools in our quasi-experimental design, 144 students completed both the pre- and post-survey on attitudes toward science (Table 4). Recall that this survey measured two dimensions of attitude: interest and perceived value. At all three schools the main effect of Time on Interest ratings was not significant, $F(1, 24) = 2.15, p = .16; F(1, 67) = .03, p = .87; F(1, 47) = .15, p = .70$ (Schools 1, 2, and 3, respectively). On average, students reported moderate levels of interest in science at pre- and post-test. At Schools 1 and 3, the interaction between Time and Group was not significant, $F(1, 24) = 1.15, p = .30; F(1, 47) = .10, p = .75$, respectively. At School 2, however, there was significant interaction between time and group, $F(1, 67) = 19.65, p < .001$, partial $\eta^2 = .23$, with CENS students reporting less interest and Comparison students reporting higher interest at post-test than at pre-test (see Figure 3).
Table 4. Number of students who completed attitude survey pre and post, by school and condition.

<table>
<thead>
<tr>
<th></th>
<th>School 1</th>
<th>School 2</th>
<th>School 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>13</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>CENS</td>
<td>13</td>
<td>41</td>
<td>25</td>
</tr>
<tr>
<td>TOTAL</td>
<td>26</td>
<td>69</td>
<td>49</td>
</tr>
</tbody>
</table>

With regard to the Perceived Value of Science, again there was no main effect of Time at any of the three schools, $F(1, 24) = .84, p = .37; F(1, 67) = 1.82, p = .18; and F(1, 47) = 2.40, p = .13$. On average, students moderately endorsed statements about the positive value of science. Nor was there an interaction between Time and Group at any of the three schools, $F(1, 24) = 2.79, p = .11; F(1, 67) = 1.63, p = .20, F(1, 47) = 1.78, p = .19$, for Schools 1, 2, and 3, respectively. In general, students’ perceived value ratings were relatively flat (See Figure 4). The exception here was that Comparison students at School 1 had lower means at post-test than at pre-test; as with Content Scores, the non-significant interaction effect found for this school is not interpretable as it may reflect a lack of statistical power or truly no difference in means.

Figure 3. Mean Interest Rating (± SE) as a function of Time and Group.
Attitudes about inquiry

Our survey to measure students' attitudes about inquiry were given to students in all five schools who participated in the CENS unit. A total of 153 students took this survey; 89 in the three schools included in the quasi-experimental design. To our knowledge, we are the first to ask students to rate their interest in specific tasks related to inquiry. We chose our four tasks to ask about (choosing data, analyzing data; writing explanations; using data to support claims) because they were salient to the work students did in this unit. Obviously, they are not the only tasks of inquiry that might be of interest to researchers.

We collapsed the four dimensions of value across task items to get an impression of students' general attitudes toward these inquiry tasks. The responses students gave at each school are shown in Table 5. Recall that students responded to each item on a 5-point Likert scale, with 5 being the highest value (i.e., most different, most interesting, etc.). Students reported that these tasks were modestly different from their usual school work, but they generally did not find them interesting except for students at School5, who reported moderate interest. Across schools, however, students reported high levels of usefulness to being asked to work with and explain
data. Finally, overall students did not perceive these tasks as especially more difficult than their typical school work.

Table 5. Means (SD) of inquiry attitudes, averaged across tasks.

<table>
<thead>
<tr>
<th></th>
<th>School1 (n=18)</th>
<th>School2 (n=44)</th>
<th>School3 (n=27)</th>
<th>School4 (n=34)</th>
<th>School5 (n=30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different</td>
<td>2.95 (1.0)</td>
<td>3.17 (.99)</td>
<td>3.12 (1.0)</td>
<td>3.32 (.98)</td>
<td>3.29 (.99)</td>
</tr>
<tr>
<td>Interesting</td>
<td>2.99(.82)</td>
<td>2.48 (.80)</td>
<td>2.23 (.84)</td>
<td>2.63 (.74)</td>
<td>3.45 (.98)</td>
</tr>
<tr>
<td>Useful</td>
<td>4.00 (.84)</td>
<td>4.33 (.71)</td>
<td>3.78 (.84)</td>
<td>3.41 (.90)</td>
<td>4.42 (.51)</td>
</tr>
<tr>
<td>Difficult</td>
<td>3.17 (.98)</td>
<td>2.52 (1.0)</td>
<td>2.84 (.10)</td>
<td>2.35 (.76)</td>
<td>2.37 (.88)</td>
</tr>
</tbody>
</table>

We ran one-way ANOVA on each variable to explore school differences, using Tukey post-hoc t-tests to see how schools may have varied from each other. There were no differences between schools in students' perceptions of the tasks being different from what they normally do in science. Students at these five schools varied significantly in their level of interest in these inquiry tasks, $F(4, 148)=7.511$, $p<.001$. Post-hoc Tukey t-tests showed that students at School5 found the tasks more interesting than students at all other schools, and students at School1 found them more interesting than students at School2, School3, and School4. We also found significant differences in reported levels of usefulness, $F(4,148)=9.811$, $p<.01$. Here the post-hoc tests showed that students at School2 and School5 were more likely to find the inquiry tasks useful than students at School3 or School4. Lastly, there were also differences between schools in perceived difficulty of these tasks, $F(4,148)=3.087$, $p<.05$, with post-hoc tests revealing that students at School1 were more likely to report the tasks being difficult than students at either School4 or School5. All other pair-wise comparisons of schools on these three variables – interest, usefulness, difficulty – were not significant.

We are in the process of analyzing the reasons that students gave for their ratings on each dimension. So far, students tend to see the usefulness of these tasks in schools terms or in vague life values. For instances, students reported it was useful to write explanations because they had
to do that in other classes or "later in life you will have to write many explanations." Our ongoing analyses will discern the kinds of reasons students give for their ratings and their depth.

**Students' ideas about scientific argumentation**

Thirty three students from School5 were interviewed individually on the essays they wrote as the culmination of the CENS unit, as described above in the Methods section. Recall that we analyzed students' responses to three key questions: Why were they certain of a claim? Why were they uncertain? What would be the best way to convince someone in science? Students gave a total of 110 warrants across these three questions. Four types of warrants emerged from our analysis. *Authority* warrants were appeals to some authority, usually the teachers. *Fact* warrants were appeals to the factual nature of a claim, without mentioning some data source. In contrast, when students appealed to some particular data we coded these as *empirical* warrants. Lastly, students justified claims through some *causal* appeal, such as citing one of the target scientific principles of the unit as a warrant for a specific claim.

This distribution of types of warrants across contexts is shown in Figure 5. Across all three warrant contexts (certain, uncertain, "global" persuasion), half of the warrants students gave were empirical, and students were much more likely to give empirical warrants than any others, $X^2(4) = 335.93, p < .001$. Students were also more likely to use empirical warrants to justify certain claims ($X^2(4) = 56.17, p < .001$), to explain their uncertainty about a claim ($X^2(4) = 32.54, p < .001$), and as the best way to convince someone else of a claim ($X^2(4) = 52.66, p < .001$). Empirical warrants for uncertain claims referred to a lack of data or inability to understand data, such as, “I did not have any data for this so I wasn’t sure," or, “I was unsure about this one [claim] because I did not understand the graph.” Such statements are not warrants for the claims; rather, they reflect the kind of warrant students would prefer to be less uncertain. One trend in
the interviews masked by these data is that half (17) of the students we interviewed reported that they were not uncertain about any of their claims.

Figure 5. Preferred warrants for certain and uncertain claims and for general persuasion.
Besides this general preference for empirical warrants, we also found that individual students had consistent preferences for types of warrants. If a student used the same warrant type in two or more contexts then this was considered their preferred warrant. If a student used a different warrant for each of the different contexts then they were considered to have no preferred warrant. If a student used two or more warrant types in two or more of the contexts then they were considered to have multiple preferred warrants. Eighty-eight percent of the students had a preferred warrant. Fifteen percent of the students had multiple preferred warrants, of these 80% had empirical as one of their preferred warrants. Those students with a single preferred warrant were more likely to prefer empirical warrants than the other types, \( \chi^2(5) = 30.09, p < .001 \).

**Teacher adaptations**

We used classroom observations to examine classroom teaching and learning practices while students were engaged in CENS vs. Comparison classrooms. Table 6 shows that, at each school, more events indicating active and inquiry-type learning took place in CENS classrooms than in Comparison classrooms. Moreover, CENS classrooms showed fewer characteristics of teacher-directed lessons than Comparison classrooms. Our unstructured observations of CENS classrooms suggested that though CENS teachers used more inquiry practices than their Comparison peers, their lessons were not as inquiry-oriented as the unit developers had intended.
Table 6. Frequency of instructional activities in CENS and comparison classrooms.

<table>
<thead>
<tr>
<th></th>
<th>Inquiry Checklist</th>
<th>Frequency of Inquiry Events</th>
<th>Proportion of Students Involved</th>
<th>Frequency of Teacher-directed Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>School 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CENS</td>
<td>6</td>
<td>2.4</td>
<td>2.8</td>
<td>2.2</td>
</tr>
<tr>
<td>Comparison</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4.7</td>
</tr>
<tr>
<td>School 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CENS</td>
<td>9</td>
<td>3.2</td>
<td>3.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Comparison</td>
<td>2.5</td>
<td>2.3</td>
<td>2.8</td>
<td>2.5</td>
</tr>
<tr>
<td>School 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CENS</td>
<td>8</td>
<td>3.2</td>
<td>3.5</td>
<td>1</td>
</tr>
<tr>
<td>Comparison</td>
<td>4</td>
<td>2.2</td>
<td>2.8</td>
<td>3.3</td>
</tr>
</tbody>
</table>

As mentioned above in the Methods section, we have begun microanalyses (Erickson, 1992) of the interactions between teachers and students in CENS classrooms in order to explore how the actors in the classroom construct their positions in relation to science. These analyses are at a nascent stage, and we are focused on how teachers' ways of talking about what students should be doing during particular tasks, as well as the broader purpose of these tasks, positions students as "students" in a classroom or as "scientists" investigating an interesting problem. This is not as dichotomous as it sounds, at least not in the data we see. In the interest of brevity, we will not present the full scope of the analyses we have done here. Instead, we contrast the ways that two teachers framed the CENS unit for their students.

Each of these teachers, Ms. H from School 1 and Mr. L from School 3, had attended our 1-day training session, and they have both taught at their schools for several years. We contrast their opening statements for this unit, literally the first things they said at the start of the unit. It will be clear that they position the students rather differently. Although this brief example is insufficient to show how differences in positioning play out and are sustained, it demonstrates how quickly teachers act to position students toward work in the science classroom.
Both teachers used their opening statements to the class to frame the upcoming unit as science-like, but in different ways. Each transitioned from their opening statements to having students read from a handout that we supplied. The following interaction took place in classroom one, after a very brief initial discussion in which the teacher alluded that students would be analyzing a photo to see “if you see any kinds of interesting things.” A photo of a hillside was projected onto the whiteboard. Several terms were defined on the whiteboard when class began.

(In the following transcripts, all names are pseudonyms. We use standard conventions, such as // to indicate interruptions.)

Ms. H: Okay. What this is //
Victor: //I see bushes
Ms. H: //is you guys are going to pretend that you are a scientist.
Jordan: alright.
Ms. H: You know how my husband is a scientist?
Several Students: Yes//
One student: //No
Ms. H: how he works at UCLA? What he does is he goes somewhere and he sees something like this and then he has to figure out what it is, what he wants to study, what he thinks is different, what’s the anomaly. Um, Veronica, what is an anomaly?
Veronica: an observation that doesn’t fit the pattern.
Ms. H: Very good
Robert: I don’t know!
Ms. H: It’s on the board! (“Anomaly” was defined on the whiteboard when class began
Robert: oh
Ms. H: yeah.
T: Okay, so my husband studies rocks, but you guys are going to pretend that you are plant scientists, so you’re going to be studying plants. Alright. So I want you--where’s your papers? Briannee, you didn’t pass out the papers at your table? Veronica, well, no//
Ray: //Where do we put our name at?//
Ms. H: //there’s no Anna? (gesturing to an empty chair)
Liza: Anna, Anna changed.
Frank: Anna Bananna.
Ms. H: Melissa, you wanna read the first paragraph?”
Ms. H explicitly frames the unit as a role-playing activity in which “you guys are gonna pretend that you are a scientist.” Contrast the terms of her introduction to the unit with the introduction offered by Mr. L. almost immediately after the bell rang.

Mr. L: [inaudible] Uh, investigators. As scientists. We’re going to be doing actual science, we’re going to be using some of the UCLA sensors that are in the Santa Monica Mountains to get some of our data and we’re gonna try to see what we can... discover. Plants happen to be our--the focus of the unit. Okay? ... Now, let me hand out this first sheet, and we’re going to read through it. Alright, who would like to read the first paragraph? Melissa, I’ll let you start since you have it, while I finish handing them out.

In his introduction to the unit, Mr. L. says that "[w]e're going to be doing actual science.” He furthers that idea by foreshadowing the investigation that students will eventually conduct with data from weather sensors. Alongside this framing of the upcoming investigation as “actual science,” he also refers to the work in very school-like terms as “the unit.” Rather than invoking the description of some particular real scientist as Ms. H. did, the “actual science” alluded to here is an imagined process that has to do with getting data and seeing “what we can discover.” Mr. L. also uses the word “we” repeatedly to position himself as co-investigator, distinct from the use of “you” and “you guys” by Ms. H. The subtle differences in the framing of the investigation as “science” demonstrated here are likely to be significant only if they are supported throughout the rest of this lesson and the unit. Our analyses of the remainder of this lesson show these differences to be sustained by each teacher. Continuing analyses will show whether or not each teacher maintains a consistent positioning stance toward students, or how they change.

**Discussion**

We break down this discussion of our results into two parts. First, we summarize what we believe to be the answers to the questions driving our study that are suggested by our data. After
that, we consider what this study has to say about the development of sustainable innovation in science education.

Was the Sensing the Environment unit effective? The answers seems to be yes, but it depends. At the three schools where we explicitly compared our unit to typical instruction on the same topics, we found that students at all schools learned something about the target concepts. The students at School3 had the lowest learning gains, and we saw no effects on learning from our CENS unit. Learning gains favored CENS students at School1 and School2, however, and it is noteworthy that these schools serve predominantly minority students. Our findings here are consistent with other studies that show that guided inquiry units seem to be particularly beneficial for these students (Songer et al., 2003; White & Frederiksen, 1998), and they support the effort to provide such ambitious instruction to historically underserved students. At the same time, our learning data suggest teacher differences that may influence our findings. For instance, the CENS teacher at School2 may simply be a better inquiry teacher than her peer: not only did her students learn more content (Figure 1), but they seemed to know much more about inquiry at the start of the unit (Figure 2).

Did Sensing the Environment change students' attitudes about science? The simple answer is no. The CENS students in School2 actually reported decreased interest in science after the unit. We hesitate to make too much from this finding, however, both because the effect was small and there is other evidence that students often do not initially like inquiry, either because it is harder than what they are used to or because it changes the rules of the game, so to speak, for school success (Tobin, Tippins, & Hook, 1995). At the same time, CENS students reported finding the inquiry tasks that they engaged in as quite useful, either for other school subjects or later in life. A limitation of our survey was that students' justifications of their ratings were brief.
Future research could productively examine students' attitudes towards these activities more deeply.

To our knowledge, however, we are the first to ask students about their attitudes towards specific kinds of inquiry tasks. Our findings here suggest a number of lines of research that could productively examine affective aspects of students' experience of inquiry in school and how these affect both student learning and the sorts of identities relative to science that the project for themselves. For instance, we were surprised that students did not perceive this unit as very different from what they usually do. This may stem from the large number of typical science class activities prior to the online investigation and that investigation's relative brevity; both features of our design that come from our goal to produce units that teachers can easily use in their classrooms. As we revise our materials, we are paying particular attention to a more explicit framing of the unit as a single, extended inquiry.

With regard to students' understanding of scientific argumentation, our data reveal that students definitely prefer empirical evidence to justify claims. On the one hand, this is not surprising given that the role of evidence in scientific theorizing is often overplayed in typical science classrooms. That is, evidence is often overly objectified (Lemke, 1990) and treated as if it is independent of theoretical ideas. Yet, the relation between theory and evidence in actual scientific is a dialectic – theory often drives the collection of data, while also being accountable to those data. Students here, as we have seen before (Sandoval & Millwood, 2005), treat evidence as objective. Still, given the importance of evidence in scientific explanation, this preference for empirical warrants is a productive idea upon which to build. One analysis we have yet to complete is whether or not students' argumentation is related to their content learning as measured on our learning test.
The most striking feature of our results is their variability. It seems clear that the variability in student outcomes results both from variations in the student populations at each school and from differences in teaching. Students at School3 knew more about the target science content before they started the unit, in both classes. This limited their possible learning gains. As we mentioned above, the strong content learning gains in the CENS class in School2, along with those students' apparently strong understanding of inquiry prior to the unit, suggests that that teacher may have been particularly good at inquiry teaching. Still, in all schools we observed more inquiry-type classroom practices in CENS than Comparison classrooms. Students in these classrooms were asking more questions, discussing ideas with their peers, and generating hypotheses about and conclusions from data. We interpret these results cautiously, to understand the contexts in which some of the observed learning occurred. Readers should bear in mind that none of the inquiry or teacher-directed practices are in and of themselves positive or negative. Our ongoing interaction analyses of the three CENS teachers for whom we have detailed records of their implementation will help to illuminate particular features of teaching practice that affect student outcomes and will provide ideas for supporting good practice through curricular materials.

We are aware that there are a number of additional analyses to be done of these data. We have followed a longstanding practice of design-based research by collecting large amounts of disparate data. The findings we have reported here represent only a first stage of analysis. The second stage should explore relations between various outcomes: between attitude and learning gains, argumentation and learning gains, argumentation and attitude, teaching practices and attitude, and so forth. A question for us, and any other research team doing this sort of work, is how much of these analyses ought we do at this stage of our project? Obviously, we should
conduct those analyses that will help us to improve our materials, both in the short term and for their long-term sustainability.

Sustainability is likely to be achieved only if curricular materials such as these can communicate to teachers the goals they can achieve and provide strategies for achieving those goals (Davis & Krajcik, 2005). We perceive our work on this project as being near the beginning of a research trajectory. At this point, we have little to say about long-term sustainability. Our immediate research aim is to make Sensing the Environment and the other modules we are developing more usable in a range of classrooms. This is related to sustainability to the extent that "usable" and "sustainable" overlap. We suggest that our study, while still young, makes two contributions to discussions of sustainability. First, we have expanded the space of outcomes of interest for inquiry-oriented interventions to include affective aspects of students' experiences rather than just what they learn about science. It matters that we add such outcomes because in the long term we want students to develop positive attitudes toward science, both to increase the number of ethnic minority students who may eventually choose science careers and to promote broad public understanding of science. Second, we have included analyses of teacher variations in implementation at the very start of our research trajectory. This will improve our ability to develop adaptable and educative curricular materials, which should in turn make it more likely that our materials are usable in a variety of school contexts.

Acknowledgements

The work reported here was supported by National Science Foundation grants NSF-CCR-0120778 and NSF-ESI-0352572, and by a NIMH pre-doctoral training fellowship to Millwood. The opinions expressed here are our own, however, and do not necessarily reflect the views of NSF or NIMH. We thank Deborah Estrin, Christine Borgman, and other CENS staff for their
support. We especially thank the teachers and students who participated in this work. Thanks also to Philip Bell for introducing us to the phrase "hothouse projects." More information about CENS can be found at cens.ucla.edu and our ongoing pre-college education work at censei.ucla.edu.
References


