Running Head: PROMISE AND PITFALLS OF INSTRUCTIONAL LOGS

Identifying and Measuring Factors Related to Student Learning: the Promise and Pitfalls of Teacher Instructional Logs

P. Sean Smith

R. Keith Esch

Horizon Research, Inc.

Author Note

This research was supported by National Science Foundation grant DUE-0335328. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this paper should be addressed to P. Sean Smith, Horizon Research, Inc., 326 Cloister Court, Chapel Hill, NC 27514. E-mail: ssmith62@horizon-research.com.

Abstract

Existing measures have produced inconsistent and weak evidence for claims about the relationships among teacher content knowledge, classroom practice, and student learning. The ATLAST project has developed pairs of assessments—for teachers and students—in three middle grades content areas: force and motion, flow of matter and energy in living systems, and plate tectonics. The project also developed a novel instructional log. This paper describes the development of these pairs of multiple-choice assessments and the log. We explain the types of teacher items included in the assessments and describe reliability and validity information for all six measures. We used these instruments and classroom instruction information gleaned from teacher-generated logs to explore the relationships among teacher content knowledge, classroom instruction, and student learning. Our findings suggest the relationships among teacher knowledge, amount of instruction, and student learning depend on the science content; and that teacher instructional logs do not provide sufficient evidence to gauge the quality of instruction.

Identifying and Measuring Factors Related to Student Learning: the Promise and Pitfalls of Teacher Instructional Logs

Discussions of science teacher quality tend to be theoretical and divisive. The lack of consensus can be attributed in part to a weak empirical basis for the competing arguments. A case in point is the debate over the value of teacher preparation programs. All sides claim empirical support, sometimes using the same studies; however, the commonly used measures of teacher characteristics tend to be far removed from what we care most about—classroom instruction and student learning.

Despite a lack of consensus on what constitutes "teacher quality," there is broad agreement that teacher knowledge of disciplinary content directly and positively affects classroom practice and, ultimately, student learning. Put plainly, teachers cannot teach well what they do not know. It is interesting to note that although this premise is logical, the empirical support is thin, largely because of a lack of suitable measures. Studies typically rely on proxies of teacher content knowledge, for example certification type (Goldhaber & Brewer, 2000), undergraduate major (Monk, 1994), and courses taken (Druva & Anderson, 1983). Few studies use direct measures of teacher content knowledge. And student measures tend to have weak psychometric properties, or they are very broad (e.g., state-administered assessments), further limiting the likelihood that relationships between teacher knowledge of particular content and student learning will be detected.

An extensive review of research on science teacher knowledge supports the claim that the field lacks appropriate measures of teacher content knowledge. Horizon Research, Inc. (HRI) conducted a comprehensive and rigorous literature review as part of its Math Science Partnership Knowledge Management and Dissemination Project (DUE-0445398). Many of the instruments that were used to measure teacher knowledge were developed by researchers for particular studies, and these rarely had established psychometric properties. Further, the measures tended to be narrowly focused on the purposes of the particular study. We argue that only with tightly aligned pairs of teacher and student measures can a clear picture of the role of content knowledge in science teaching emerge.

Even more closely related to student learning is what happens in the classroom. Measuring classroom instruction, however, presents logistical and methodological challenges. Observing instruction, particularly in large-scale, geographically dispersed schools, is prohibitively expensive. Researchers have explored instructional logs as a proxy for direct observation with mixed results (e.g., Rowan, Camburn, & Correnti, 2004; Rowan, Harrison, & Hayes, 2004).

In this paper, we describe the development of assessment measures, an instructional log, and an associated analysis protocol used in a study of factors associated with student achievement. We focus on the affordances and limitations of the instructional log.

Instrument Development

Development of the Assessments

The ATLAST (Assessing Teacher Learning About Science Teaching) assessment development process is depicted graphically in Figure 1. Each component is described in the text that follows.

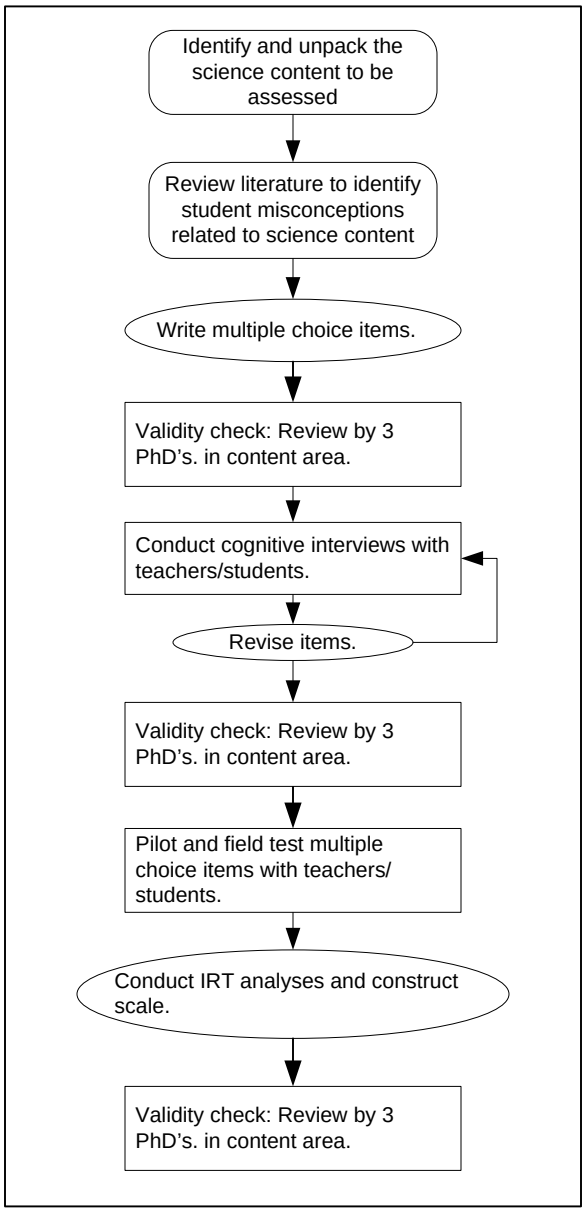


Figure 1. The ATLAST Assessment Development Process

For the ATLAST project, HRI selected three relatively narrow slices of middle grades science content. The original content statements were taken directly from *Benchmarks for Science Literacy* (AAAS, 1993). Each topic is central to its discipline (life, physical, and earth science), and each appears in virtually every state standards document.

1. Flow of matter and energy in living systems: Food provides the molecules that serve as fuel and building material for all organisms. Plants use the energy from light to make sugars from carbon dioxide and water. This food can be used immediately or stored for later use. Organisms that eat plants break down the plant structures to produce the materials and energy they need to survive. Then they are consumed by other organisms. (AAAS, 1993, p. 120)
2. Force and motion: An unbalanced force acting on an object changes its speed or direction of motion, or both. (AAAS, 1993, p. 90)
3. Plate tectonics: The solid crust of the earth—including both the continents and the ocean basins—consists of separate plates that ride on a denser, hot, gradually deformable layer of the earth. The crust sections move very slowly, pressing against one another in some places, pulling apart in other places. Ocean-floor plates may slide under continental plates, sinking deep into the earth. The surface layers of these plates may fold, forming mountain ranges. (AAAS, 1993, p. 74)

These statements were then unpacked into discrete, assessable statements, or "sub-ideas," that provided guidance to item writers and ensured that the items assessed subtleties of the content hidden in the general statement. In addition to unpacking the topic, we reviewed the literature on student/adult thinking related to the content. We associated the misconceptions (or prior/naïve conceptions) with the sub-ideas, providing further guidance to item writers. In particular, the misconceptions often provided useful distractors for multiple-choice items, both for teachers and students.

HRI staff developed items in teams organized by content expertise. Once a substantial pool of items had been written, we initiated a series of team meetings to edit the items

collaboratively. The principles that guided our item writing work are described elsewhere (Taylor and Smith, 2009), but two criteria carried the most weight for ensuring item validity—necessity and sufficiency (Stern & Ahlgren, 2002). An item met the necessity criterion if it required that the respondent know the target content in order to answer the question correctly. An item met the sufficiency criterion when the knowledge in the target content was all the respondent needed to know to answer the item correctly; answering the item correctly required no science content knowledge outside the target content. When editing of the initial pool of items was complete, the items were sent to three content experts to be reviewed for accuracy.

Once expert feedback had been incorporated in the items, we initiated several rounds of cognitive interviews with teachers (for the teacher assessment items) and students (for the student assessment items). The point of these interviews was to maximize item validity. The interviews revealed whether (a) teachers/students answered the question we intended them to answer (or did they interpret it so differently than we intended that they answered a different question?); and (b) teachers/students used their knowledge of the targeted content to answer the question. For each content area, a pool of 60 – 80 multiple-choice items was then piloted with several hundred teachers or students, depending on the audience. Data from the teacher and student pilots were analyzed using an item response theory (IRT) framework. Results from the analysis were used to reduce the item pool to between 30 and 40 for each content area and audience (teacher or student). Items were selected using two criteria: adequacy of measurement properties and coverage of the content domain for the assessment. This smaller pool of items was then field tested with several hundred teachers/students. Using results of the field test analysis, we selected approximately 30 items for each of the assessments, again using the criteria of measurement property adequacy and content domain coverage.

Design of the Teacher Assessments

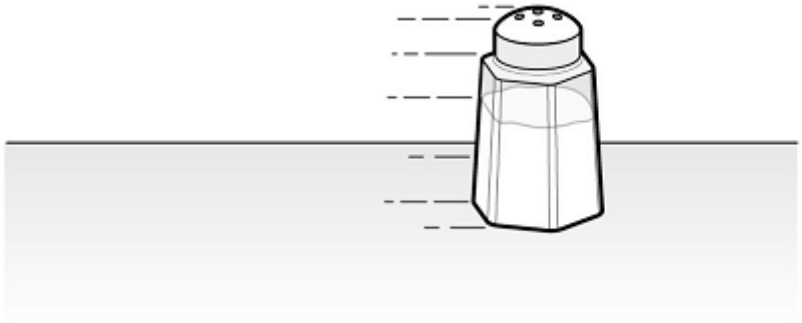
The primary focus of the teacher assessments is content knowledge, which, as a construct is distinct from other domains of teacher knowledge. For example, teachers vary not only in their knowledge of disciplinary content but also in their knowledge of how students think about science concepts and in their knowledge of effective strategies for engaging students with science content and making sense of it. Both of these examples fall in the realm of knowledge that Shulman (1986) called "pedagogical content knowledge." The literature on pedagogical content knowledge in science (e.g., Carlsen, 1999; Magnusson, Krajcik, & Borko, 1999; Veal and MaKinster, 1999; Wilson & Berne, 1999), suggests multiple *content-specific* domains of teaching knowledge. Ultimately, we constrained our assessment items to teacher knowledge of disciplinary content for two reasons. First, disciplinary content knowledge is the foundation on which all other content-specific domains of knowledge are built. Second, it has proven to be the most measurable domain of teacher knowledge to date. We have developed three types of items for assessing teacher content knowledge, each set in instructional contexts: (a) assessing knowledge of science content; (b) assessing content knowledge through the analysis of student thinking; and (c) assessing content knowledge through instructional decision-making.

Examples of our three types of teacher assessment items are shown in Figures 2-4. The correct answer in each item appears in bold text.

A teacher gives her students the following question on an end-of-unit test.

Student Assessment Item

A boy slides a saltshaker along a table toward the right. As the saltshaker slides, in which direction does the force of friction act on the saltshaker?



What would be the correct answer?

A. To the right
B. To the left
C. Upward
D. Downward

Figure 2. Example of an item assessing teacher knowledge of science content.

In a lesson on plant growth, a teacher is discussing plants' use of light energy from the Sun. During the discussion, one student says, **"Plants need the light to grow, but they don't change the light at all. It's like when you're reading a book, and you need the light to help you read."**

Which one of the following ideas about the role of light energy in photosynthesis does the student seem to be missing?

A. Light energy is changed into sugars in the process of photosynthesis.
B. Light energy is changed into another form of energy in the process of photosynthesis.
C. Light energy is the energy source for the process of photosynthesis.
D. None. The student seems to have an accurate understanding of the role of light energy in the process of photosynthesis.

Figure 3. Example of an item assessing teachers' content knowledge through the analysis of student thinking.

In a class discussion, a teacher asks his students to describe Earth's plates. One student says, **"There are thousands of plates that are moving and causing changes to Earth's surface."**

Based on this statement, which one of the following should the teacher do next to further this student's understanding of Earth's plates?

A. Discuss the types of geological features plate movement can cause.
B. Have students outline the boundaries of the plates on a map.
C. Introduce students to the specific ways in which plates move.
D. Demonstrate how the plates move as a result of convection.

Figure 4. Example of an item assessing teachers' content knowledge through instructional decision-making.

Note that in Figure 3, answer choice C includes a scientifically correct statement. However, the choice is incorrect because it does not relate to the student thinking represented in the question. To answer this type of question, teachers must process the science content and the

student thinking, a task they perform daily in the classroom. As such, this type of item simulates the work teachers do with content knowledge. Similarly, in Figure 4, each of the answer choices presents an instructional activity that is reasonable to include in a unit on plate tectonics. Only one choice, though, pertains to the student comment in the question. To reason through this type of item, a teacher must process the science content and the student thinking, then evaluate each of the choices in light of both. Again, this type of item represents one of the ways teachers draw on content knowledge in their work.

Because of the instructional contexts and the multi-step reasoning required by items shown in Figures 3 and 4, we believe these items are beginning to address important teacher knowledge beyond disciplinary content knowledge. The items do not separate as a distinct factor in dimensionality analysis, but our interviews suggest the items are quite challenging and require teachers to use their knowledge in ways that are authentic to their profession.

The student items we have developed in ATLAST are much more straightforward in that they do not include instructional contexts. An example is shown in Figure 5.

Which one of the following diagrams shows all the ***horizontal*** forces acting on a cart once it leaves your hand and is rolling across a table. (Correct answer: A)

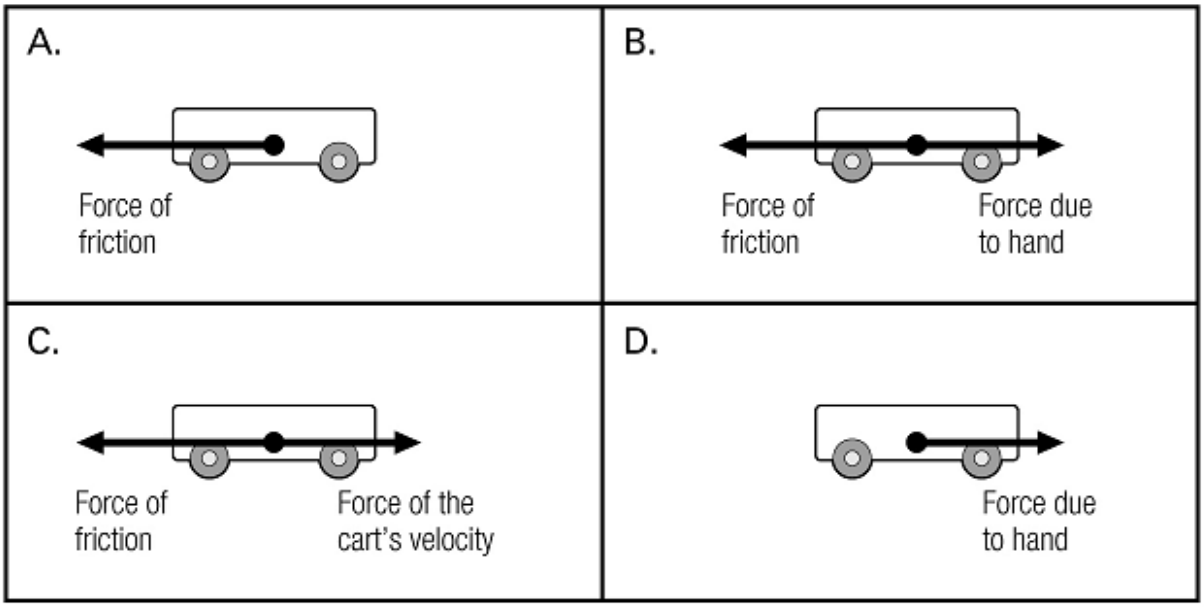


Figure 5. Example of a student assessment item.

Validity of the Assessments

Several lines of evidence support the validity of the ATLAST teacher and student assessments. First, each question was evaluated through cognitive interviews, revealing whether teachers and students use the targeted content knowledge to answer the questions. Second, after piloting and field testing, each question was also evaluated quantitatively with regard to its discriminatory power. Third, we conducted dimensionality analyses, including factor analysis and cluster analysis. The vast majority of items loaded on a single dominant factor (e.g., understanding of force and motion). Those that did not were discarded. Finally, content experts reviewed each of the final assessments and indicated that the measures adequately covered the content domain.

Reliability of the Assessments

Reliability data for the assessments are summarized in Table 1. Note that each of the internal reliabilities is .75 or higher. Test-retest reliability data are available for only the teacher measures. These were established through a study in which approximately 100 teachers completed the same assessment two weeks apart with no intervening instruction. A different sample of teachers was used for each of the three assessments.

Table 1.

Reliabilities of Student and Teacher Assessments

Measure	IRT Reliability	Test-retest Reliability
Flow of Matter and Energy in Living Systems		
Student	.78	n/a
Teacher	.85	.93
Force and Motion		
Student	.75	n/a
Teacher	.85	.88
Plate Tectonics		
Student	.86	n/a
Teacher	.86	.94

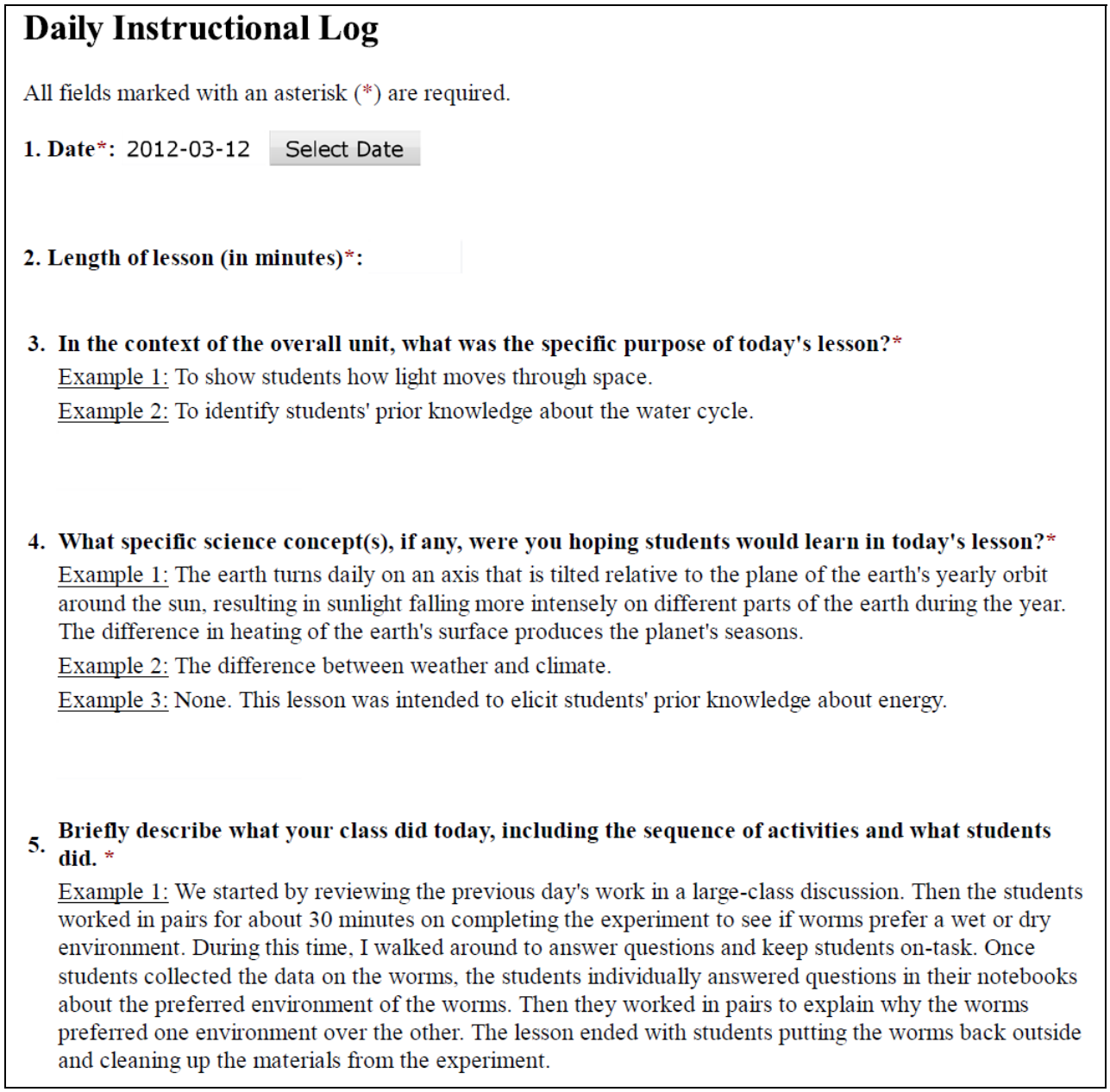
Development of the Instructional Log

In conceptualizing the instructional log, we had a number of goals in mind. First, we wanted the log to serve as a proxy for classroom observation, which meant gathering enough detail to make judgments about student opportunity to learn. A second, competing goal was to

develop a log that respected teachers' time, requiring as little effort to complete as possible. We set 15 minutes of completion time per lesson as a design constraint. Finally, we wanted the log to be convenient for teachers to complete and for researchers to analyze, a goal we accomplished by situating the log in a web-based environment.

Eliciting Detail

The log requested narrative descriptions of each day of instruction in the targeted unit. To make this task manageable for teachers, we broke the information into pieces and provided sample responses, as shown in Figure 6.



Daily Instructional Log

All fields marked with an asterisk (*) are required.

1. Date*: 2012-03-12 Select Date

2. Length of lesson (in minutes)*:

3. In the context of the overall unit, what was the specific purpose of today's lesson?*

Example 1: To show students how light moves through space.

Example 2: To identify students' prior knowledge about the water cycle.

4. What specific science concept(s), if any, were you hoping students would learn in today's lesson?*

Example 1: The earth turns daily on an axis that is tilted relative to the plane of the earth's yearly orbit around the sun, resulting in sunlight falling more intensely on different parts of the earth during the year. The difference in heating of the earth's surface produces the planet's seasons.

Example 2: The difference between weather and climate.

Example 3: None. This lesson was intended to elicit students' prior knowledge about energy.

5. Briefly describe what your class did today, including the sequence of activities and what students did. *

Example 1: We started by reviewing the previous day's work in a large-class discussion. Then the students worked in pairs for about 30 minutes on completing the experiment to see if worms prefer a wet or dry environment. During this time, I walked around to answer questions and keep students on-task. Once students collected the data on the worms, the students individually answered questions in their notebooks about the preferred environment of the worms. Then they worked in pairs to explain why the worms preferred one environment over the other. The lesson ended with students putting the worms back outside and cleaning up the materials from the experiment.

Figure 6. Sample page from instructional log.

Respecting Teachers' Time

The log incorporated several features designed to minimize completion time. The web-based format allowed teachers to type, rather than handwrite their responses. It remembered responses to previous questions and used them both to fill in information for teachers and to structure follow-up questions. For instance, prior to their first log, teachers entered the name of

their textbook, the publisher, and the relevant chapter(s). This information was automatically filled in for subsequent questions.

In addition to the narrative responses requested, the log presented a checklist of possible activities, shown in Figure 7.

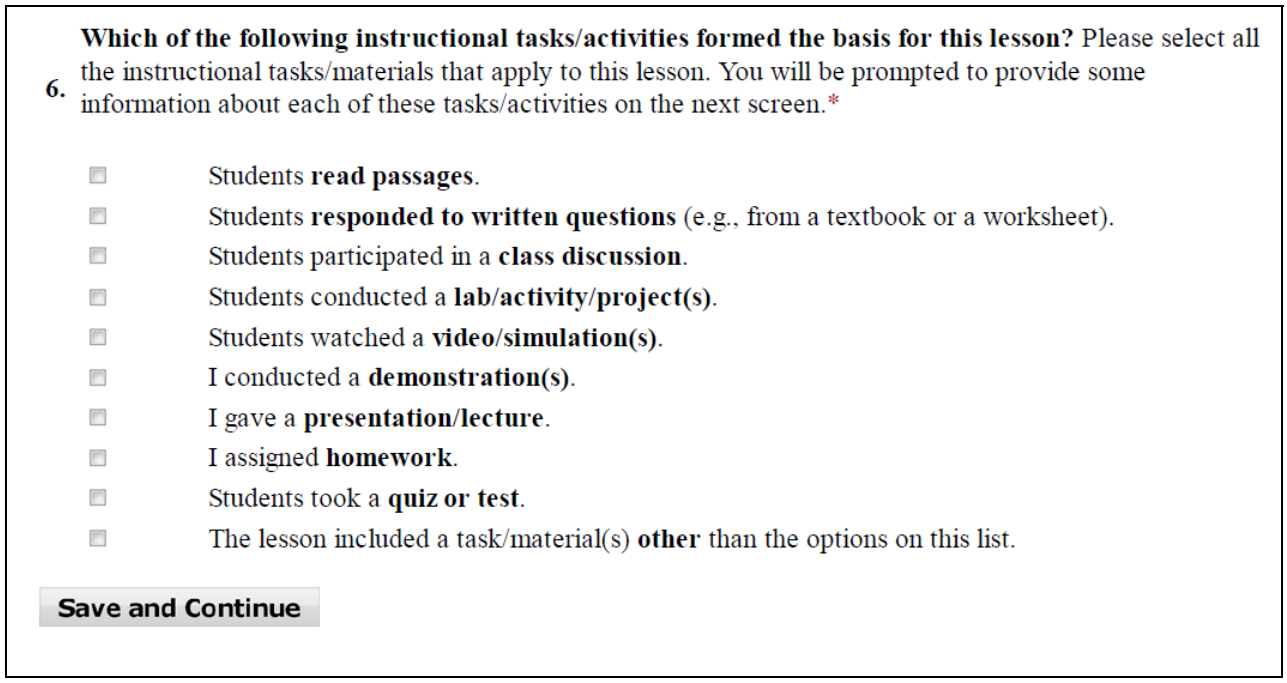
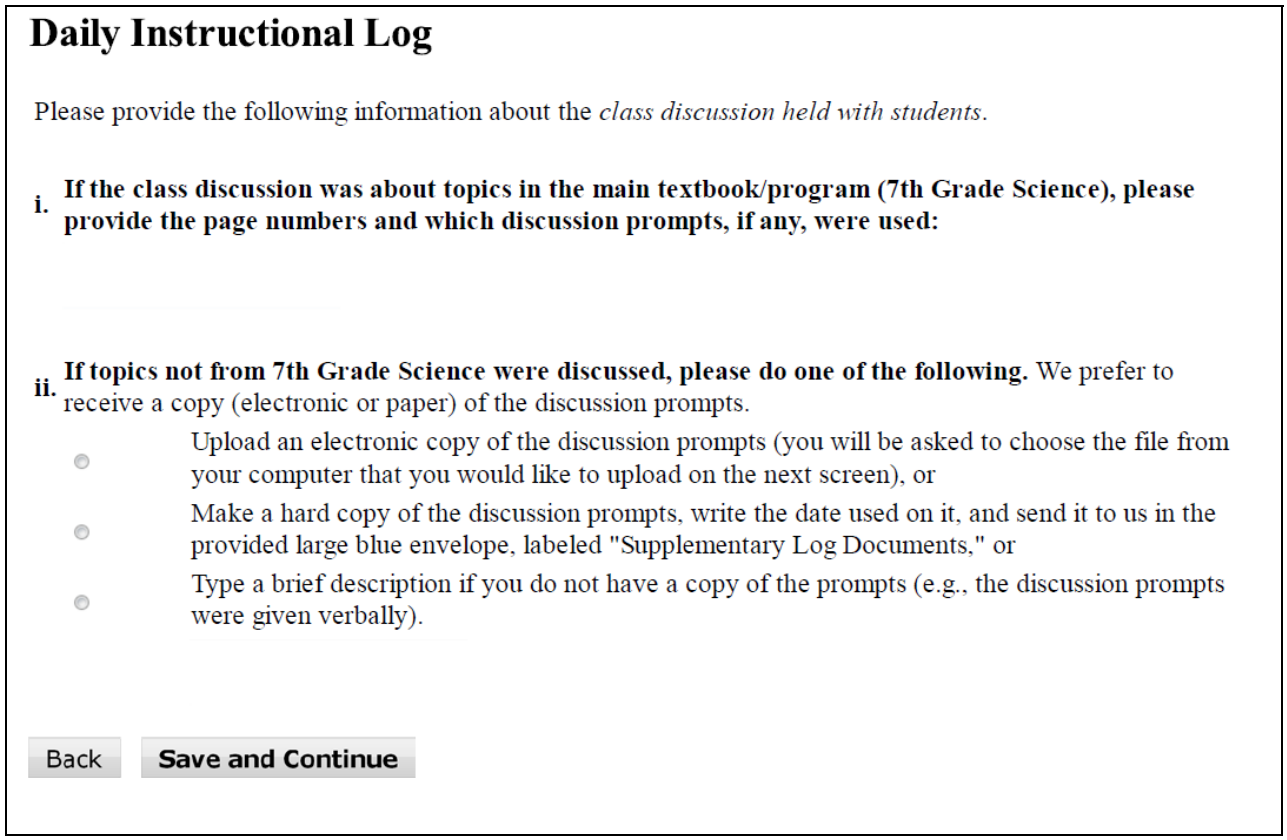


Figure 7. Checklist of instructional activities.

The questions that followed this checklist were determined by the activities selected. As such, teachers saw only the questions that were relevant to the lesson they were describing, another effort to respect teachers' time and not overwhelm them. Follow-up questions for three sample activities are shown in Figures 8 – 10. Note that in each example, teachers are given three options for providing additional detail about the activity.



Daily Instructional Log

Please provide the following information about the *class discussion held with students*.

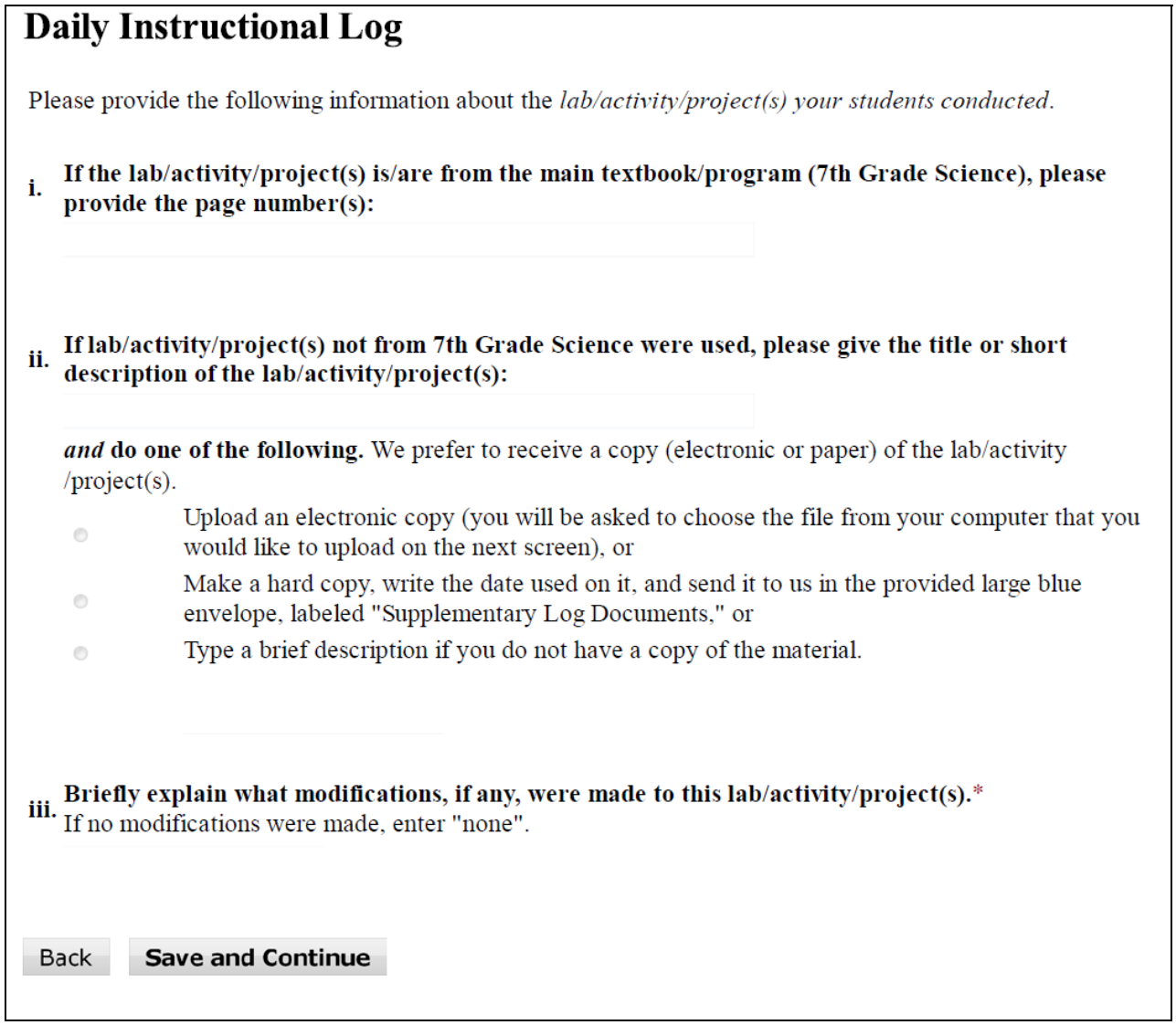
i. **If the class discussion was about topics in the main textbook/program (7th Grade Science), please provide the page numbers and which discussion prompts, if any, were used:**

ii. **If topics not from 7th Grade Science were discussed, please do one of the following.** We prefer to receive a copy (electronic or paper) of the discussion prompts.

- Upload an electronic copy of the discussion prompts (you will be asked to choose the file from your computer that you would like to upload on the next screen), or
- Make a hard copy of the discussion prompts, write the date used on it, and send it to us in the provided large blue envelope, labeled "Supplementary Log Documents," or
- Type a brief description if you do not have a copy of the prompts (e.g., the discussion prompts were given verbally).

Back **Save and Continue**

Figure 8. Follow-up questions for class discussion.



Daily Instructional Log

Please provide the following information about the *lab/activity/project(s) your students conducted.*

i. **If the lab/activity/project(s) is/are from the main textbook/program (7th Grade Science), please provide the page number(s):**

ii. **If lab/activity/project(s) not from 7th Grade Science were used, please give the title or short description of the lab/activity/project(s):**

***and* do one of the following.** We prefer to receive a copy (electronic or paper) of the lab/activity /project(s).

- Upload an electronic copy (you will be asked to choose the file from your computer that you would like to upload on the next screen), or
- Make a hard copy, write the date used on it, and send it to us in the provided large blue envelope, labeled "Supplementary Log Documents," or
- Type a brief description if you do not have a copy of the material.

iii. **Briefly explain what modifications, if any, were made to this lab/activity/project(s).***
If no modifications were made, enter "none".

Back **Save and Continue**

Figure 9. Follow-up questions for lab/activity/project.

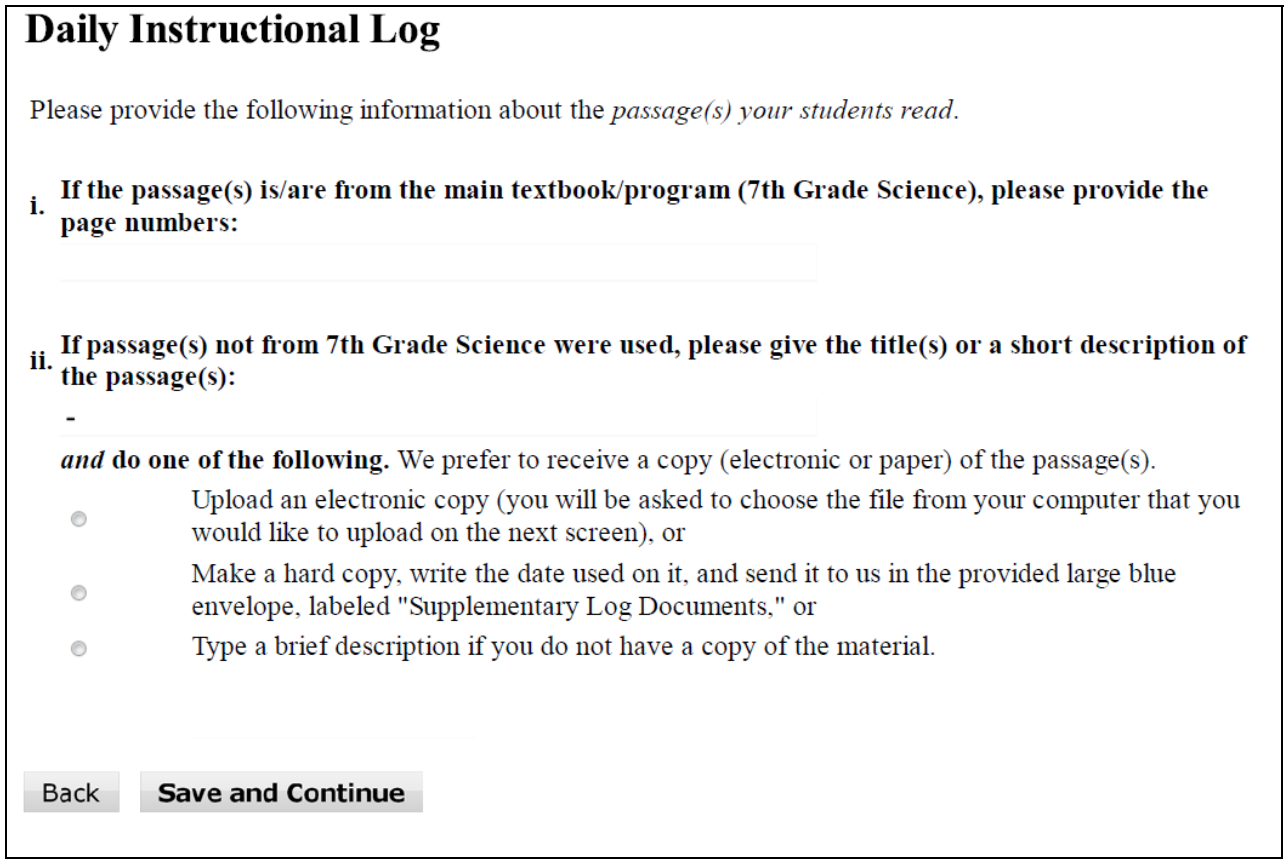


Figure 10. Follow-up questions for reading.

Convenience for Teachers and Researchers

The web-based, dynamic format maximized convenience for teachers. It also reduced burden on the researchers. By structuring the log in a series of discrete questions, we ensured that all responses took the same format. The data were already sorted to some extent, which shortened the time for analysis substantially. In addition, since teachers typed their responses, there was no need for transcription.

Development of the Analysis Protocol

Development of the analysis protocol, which spanned several years, is described briefly here. Our goal was to arrive at a measure of student opportunity to learn, defined narrowly as the opportunity provided during classroom instruction. As such, it seemed appropriate to situate our

approach to analyzing the instructional logs in a theory of teaching for understanding. Consistent with this theory, we adopted a variation of the learning cycle as the framework for measuring student opportunity to learn (Banilower, Cohen, Pasley, & Weiss, 2010):

- Situating the learning;
- Students expressing their initial ideas;
- Students examining relevant phenomena;
- Students making sense of phenomena; and
- Students making sense of the targeted idea(s).

We spent substantial time over two years identifying the phenomena that support students' in developing understanding of the benchmarks.[1] These phenomena and the associated sense-making informed the development of the analysis protocol. As we began thinking about the relevant phenomena in the three content areas, we were struck by the fact that few, if any, in plate tectonics and flow of matter and energy were directly observable. In plate tectonics, the phenomena are too big (tectonic plates, earthquakes), too slow (plate movement), or too far away (internal Earth processes) to be observed directly. In flow of matter and energy, the phenomena are essentially molecular (photosynthesis and cellular respiration), too small and too fast to be observed directly. We had a conceptual breakthrough when we made a distinction between evidentiary phenomena and primary phenomena. Evidentiary phenomena are observable, naturally occurring events that provide evidence for the primary phenomena (e.g., bubbles appear on the leaves of a plant submerged in water, providing evidence that plants produce oxygen during photosynthesis).

[1] We define phenomena as naturally occurring events embedded in or suggested by the sub-ideas (e.g., plates split and join; plants make their own food; objects move faster and faster when a constant net force is applied in the direction of motion). Phenomena are not instructional activities or data, and they are not necessarily observable.

Once we identified the evidentiary and primary phenomena for the three content areas, our attention turned to operationalizing what it meant for students to engage with phenomena, ultimately resulting in the analysis protocol, a part of which is shown in the Appendix. Reading from left to right, the first column shows the relevant primary phenomenon, but it is shaded to convey that the primary phenomenon is not actually considered in this part of the form. We included it only to give the rater a sense of where s/he was in the larger content domain. The second column shows the evidentiary phenomenon—the focus of this part of the analysis protocol—and asks the rater to provide a global sense of how the evidentiary phenomenon was addressed. In this section, "confirmatory" means that the enacted curriculum essentially "told students the answer" without giving them an opportunity to examine any evidence or do any sense-making first. "Exploratory" means that instruction asked students to engage with some form of data in order to make sense of the evidentiary phenomenon.

The rest of the table is a matrix that captures two important dimensions of instruction simultaneously: (1) how instruction presents data to students (e.g., not at all, provides data to students, students collect data), and (2) whether and how any sense-making is done from the data to the evidentiary phenomenon. The numbers in the cells represent hypothesized relative amounts of opportunity to learn about the evidentiary phenomenon. Some cells are shaded, indicating that no rating is possible. Also, the same rating appears in several cells, suggesting that students may have the same amount of opportunity in more than one way. For instance, instruction that provides substantial explanation of a phenomenon but without any data receives a rating of 2, whether no data are presented or students collect data without opportunity for sense-making. The matrix also tries to take student engagement into account. For instance,

when instruction includes sense-making, it receives a higher rating if students are involved in collecting the data than if the data are provided to them.

The table in the Appendix represents an approach to rating one instance of one evidentiary phenomenon. Our ultimate goal, however, was to rate student opportunity to learn a benchmark; that is, coherent set of ideas. Each idea in the benchmark has multiple primary phenomena; and each primary phenomenon may have multiple evidentiary phenomena. Our challenge was to develop a way to aggregate ratings for each evidentiary phenomenon to a single rating that represented opportunity to learn the set of ideas. Aggregating across so many levels introduces multiple sources of unreliability among raters (and even within a rater). We adopted two principles to maximize reliability. First, the protocol asks for ratings initially at the smallest grain size possible. Second, and closely related, the protocol requires the rater to infer as little as possible. Using this approach, ratings for evidentiary phenomena (the smallest grain size) inform ratings for primary phenomena, which inform ratings for opportunity to learn the idea, which finally inform ratings for opportunity to learn the set of ideas (e.g., ideas about plate tectonics).

Description of the Study

We designed a study to look at the relationships among teacher content knowledge, classroom instruction, and student learning of science content. We recruited approximately 200 teachers nationally for each of two studies: one on force and motion and one on plate tectonics. To be eligible for the study, each teacher had to affirm that he or she taught a unit on the content of the study as part of their normal instruction. Teachers were asked not to alter their instruction in any way.

Once enrolled in the study, each teacher (a) completed the relevant teacher assessment shortly before the unit of instruction, (b) administered the relevant student assessment just before

and just after the unit; and (c) completed a daily web-based instructional log during the unit. Teachers completed their assessment on-line. Teachers administered a paper-and-pencil assessment before and after the unit of instruction. Teachers also completed the web-based instructional log for each day of the unit. Data collection took place from January to June 2009. There was substantial attrition at the teacher level from both studies, which we attribute to the burden associated with the teacher log, despite our efforts to minimize time required for completion. Of the teachers in the Force and Motion study, 79 completed all components, compared to 107 in the Plate Tectonics study.

Results

Data from the study were analyzed progressively, beginning with the most efficient quantitative analysis. Initially, the teacher logs were mined only for amount of instructional time devoted to the topic. IRT scores (theta scores) were calculated for each student (pre- and post-unit) and teacher. Means and standard deviations are shown in Table 2.

Table 2.

Mean IRT (Theta) Scores on Student and Teacher Assessments

Measure	Pre-test		Post-test	
Force and Motion	Mean	S.D.	Mean	S.D.
Student (n=1689)	-0.003	0.640	0.496	0.776
Teacher (n=79)	0.248	0.823	n/a	
Plate Tectonics				
Student (n=2261)	0.139	0.851	0.862	0.990
Teacher (n=107)	0.200	0.833	n/a	

The mean minutes of aligned instruction for all teachers in each study is shown in Table 3.

Table 3.

Minutes of Instruction

Content Area	Minutes	S.D.
Force and Motion (n=79)	180.50	97.68
Plate Tectonics (n=107)	208.29	116.73

Data were analyzed using hierarchical linear modeling (HLM) to account for the variation shared by students in the same class. Student pre- and post-test scores were entered at the first level. Student demographic variables (race/ethnicity, gender, grade in school, English language learner status) were entered at the second level. Teacher variables (assessment score and hours of instruction) were entered at the third level.

In the Force and Motion study, amount of instruction by itself did not predict student learning. Teacher knowledge was a significant but weak predictor of student learning, both by itself and with amount of instruction in the model. An increase of one standard deviation in teacher score translated to an increase of 0.12 standard deviations in student learning above the average gain. In Plate Tectonics, neither amount of instruction nor teacher knowledge predicted student learning, either by themselves or in combination.

These findings suggest that the relationships among teacher knowledge, amount of instruction, and student learning depend on the content students are studying. Most surprising are the findings that (1) amount of instruction by itself does not predict student learning of force and motion concepts, (2) the weak or non-existent relationship between teacher content knowledge

and student learning; and (3) neither teacher knowledge nor amount of instruction predict student learning in plate tectonics. These findings led us to consider the instructional logs in more detail. Our goal was to extract from these logs a measure of student opportunity to learn, which we could then add to the analysis model.

Using the instructional logs and artifacts from a handful of teachers, we used the analysis protocol, a portion of which is shown in the Appendix, and achieved high inter-rater reliability across several raters. We intentionally chose teachers who provided substantial detail in their logs. However, as we began applying the protocol to other teachers' materials, our inter-rater reliability decreased. We revised the protocol in ways that we thought would reduce the amount of inference required of raters and raise reliability, but without success. We traced the source of our disagreement to ratings related to sense-making. We found that the logs, while helpful, typically did not include the rich description needed to rate students' opportunities to make sense of ideas. Without actually seeing what happened in the classroom, we did not feel confident in rating these opportunities.

In addition to unreliability (a fatal flaw in itself), our protocol had the disadvantage of being labor intensive. We estimated that for a typical instructional unit (two to three weeks in duration), a rater would need two days to analyze the teacher logs and instructional materials in order to arrive at a rating. A moderately large quantitative study of 100 teachers would therefore require close to one year of analyst time.

We pursued a different approach that reduced both the amount of inference and the amount of time required of raters. Each idea, primary phenomenon, and evidentiary phenomenon was rated on a three-point scale where –1 was "addressed, but inaccurately"; 0, "not addressed"; and 1, "addressed accurately." Although we forfeited much of the detail that we thought would

predict student learning, inter-rater reliability with the new approach improved substantially, and we used it to rate opportunity to learn for 30 plate tectonics units, each taught by a different teacher. We selected the units purposively, ensuring variation in class mean gain scores (the difference between the post-unit and pre-unit score) on the ATLAST student plate tectonics assessment. Our analysis, however, revealed that there was no relationship between gain scores and the opportunity-to-learn rating for the unit, suggesting that although the new approach was reliable, it was not valid.

Discussion

Our attempts to measure student opportunity to learn fell short of our goals. In the process, we became convinced that some aspects of instruction, in particular sense-making, cannot be rated validly or reliably without observing instruction (either in person or through video). Even very detailed logs do not provide enough description to rate students' opportunities to engage with ideas meaningfully. Teacher self-report data are useful for many purposes, but they are limited in inferences that can be drawn about subtle and important aspects of instruction (Mayer, 1999). Although we eventually developed a measure that had good inter-rater reliability, the ratings had no relationship with a measure of student learning.

Work on the student opportunity-to-learn instruments taught us a great deal. Early in these efforts, we debated whether we should measure quality of instruction or opportunity to learn. The difference, if there is one, is subtle. But we found that when we thought in terms of quality of instruction, we tended to think about features of instruction (e.g., teacher questioning, intellectual engagement). When we thought in terms of opportunity to learn, we focused more on the science content and the specific opportunities students have to engage with content and make sense of it. For our purposes, the latter approach seemed most appropriate.

As with other development efforts in ATLAST, the opportunity-to-learn work impressed upon us just how different our three content areas are. We began the work in force and motion, where the phenomena of interest are, for the most part, directly observable. As we moved to work in plate tectonics and flow of matter and energy, we learned that student opportunity to learn is difficult to define when the most relevant phenomena are not accessible to students. Sometimes the phenomena are too small, sometimes too fast or too far away for students to observe. We subsequently altered our approach to opportunity to learn, focusing on evidentiary phenomena (observable phenomena that provide evidence of inaccessible ones) and the reasoning (or sense-making) necessary to link these phenomena to the relevant ideas. We found this distinction between types of phenomena quite helpful in other areas of our work.

References

Carlsen, W. (1999). Domains of teacher knowledge. In J. Gess-Newsome & N. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 133–144). Norwell, MA: Kluwer Academic Publishers.

Druva, C. A., & Anderson, R. D. (1983). Science Teacher Characteristics by Teacher Behavior and by Student Outcome: A Meta-Analysis of Research. *Journal of Research in Science Teaching, 20*(5), 467–79.

Goldhaber, D. D., & Brewer, D. J. (2000). Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129–45.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 95–132). Norwell, MA: Kluwer Academic Publishers.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29.

Monk, D. H. (1994). Subject Area Preparation of Secondary Mathematics and Science Teachers and Student Achievement. *Economics of Education Review, 13*(2), 125–45.

Project 2061 (American Association for the Advancement of Science). (1993). *Benchmarks for Science Literacy*. New York: Oxford University Press.

Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in 3rd grade classrooms. *Elementary School Journal*, *105*, 75–102.

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *The Elementary School Journal, 105*(1), 103–127.

Stern, L., & Ahlgren, A. (2002). Analysis of Students' Assessments in Middle School Curriculum Materials: Aiming Precisely at Benchmarks and Standards. *Journal of Research in Science Teaching, 39*(9), 889–910.

Veal, W. R., & MaKinster, J. G. (1999). Pedagogical Content Knowledge Taxonomies. *Electronic Journal of Science Education, 3*(4).

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. *Review of research in education,* 173–209.

Appendix

III. Students Examine Relevant Evidentiary Phenomena

Instances of Evidentiary Phenomenon in Enacted Curriculum

Instance of Evidentiary Phenomenon

Primary Phenomenon	Evidentiary Phenomenon	How are relevant data presented? †	How is the evidentiary phenomenon addressed?						
			Not addressed	Without data‡		With sense-making between the data and phenomenon			
				incidental	substantive	done inappropriately	*by* students	*for* students	*with* students
1. Earth's plates move.	1.1 Chains of progressively older volcanoes are formed by hot spots.	Data not presented	0	1	2				
		Data are provided to students, but are NOT developmentally appropriate for students	0	1	2	0			
	How was the evidentiary phenomenon addressed? ☐Not at all ☐Confirmatory ☐Exploratory	Developmentally appropriate data are provided to students	1	1	2	1	3	4	5
		Data are collected by students	1	1	2	1	4	5	6

† Examples of data include:

- Table or narrative description of age (absolute or relative) of island vs. distance from "fixed point" (e.g. hot spot).
- Map of island chain and hot spot with age (absolute or relative) of islands superimposed on map.

‡ "Without data" can have two meanings. First, the evidentiary phenomenon could have been addressed without students interacting with any data. Second, it could mean that the evidentiary phenomenon was addressed, AND students interacted with data, but the data were not used in sense-making for the evidentiary phenomenon; therefore, effectively, the evidentiary phenomenon was addressed without data.