CRESST REPORT 733

Maria Araceli Ruiz-Primo
Min Li
Shin-Ping Tsai
Julie Schneider

TESTING ONE PREMISE OF SCIENTIFIC INQUIRY IN SCIENCE CLASSROOMS: A STUDY THAT EXAMINES STUDENTS' SCIENTIFIC EXPLANATIONS

FEBRUARY 2008

The National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Sciences
UCLA | University of California, Los Angeles

Testing One Premise of Scientific Inquiry in Science Classrooms: A Study That Examines Students' Scientific Explanations

CRESST Report 733

Maria Araceli Ruiz-Primo[1], [4], Min Li[2], 4, Shin-Ping Tsai[2], & Julie Schneider[3]
[1]University of Colorado at Denver
[2]University of Washington
[3]University of Colorado at Boulder
[4]Stanford Education Assessment Laboratory

February 2008

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Copyright © 2008 The Regents of the University of California

The work reported herein was supported under the National Research and Development Centers, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

The findings and opinions expressed in this report are those of the author(s) and do not necessarily reflect the positions or policies of the National Research and Development Centers, the Institute of Education Sciences (IES), or the U.S. Department of Education.

TESTING ONE PREMISE OF SCIENTIFIC INQUIRY IN SCIENCE CLASSROOMS: A STUDY THAT EXAMINES STUDENTS' SCIENTIFIC EXPLANATIONS

Maria Araceli Ruiz-Primo[1, 4], Min Li[2, 4], Shin-Ping Tsai[2], & Julie Schneider[3]
[1]University of Colorado at Denver
[2]University of Washington
[3]University of Colorado at Boulder
[4]Stanford Education Assessment Laboratory

Abstract

In this study we analyze the quality of students' written scientific explanations in eight science inquiry-based middle-school classrooms and explore the link between the quality of students' scientific explanations and their students' performance. We analyzed explanations based on three components: claim, evidence to support it, and a reasoning that justifies the link between the claim and evidence. Quality of explanations was linked with students' performance in different types of assessments focusing on the content of the science unit studied. To identify critical features related with high quality explanations we also analyzed the characteristics of the instructional prompts that teachers used. Results indicated that: (a) Students' written explanations can be reliably scored with the proposed approach. (b) The instructional practice of constructing explanations has not been widely implemented despite its significance in the context of inquiry-based science instruction. (c) Overall, a low percentage of students (18%) provided explanations with the three expected components. The majority (40%) of the "explanations" found were presented as claims without any supporting data or reasoning. (d) The magnitude of the correlations between students' quality of explanations and their performance, all positive but of varied magnitude according to the type of assessment, indicate that engaging students in the construction of high quality explanations might be related to higher levels of student performance. The opportunities to construct explanations, however, seem to be limited. We report some general characteristics of instructional prompts that showed higher quality of written explanations.

Introduction

The general premise of scientific inquiry teaching is to engage students in the activities and thinking processes of scientists to develop understanding of important concepts, principles, and methods of science (National Research Council [NRC], 1996). It is important to recognize that scientific inquiry goes beyond designing experiments and/or executing procedures, using instruments, recording data, or constructing graphs. It involves understanding where scientific theories, principles, and concepts come from within a field of

study. For any activity to be counted as inquiry, there should be an *epistemic goal*: "the pursuit of a particular kind of 'account' of the world" (Sandoval, 2005, p. 1). Therefore, one fundamental activity in scientific inquiry is the construction of explanations.

The science standards documents (NRC, 1996, 2000) emphasize scientific explanations as an *essential feature*, a *fundamental ability*, and a *fundamental understanding* of scientific inquiry. In the words of the NRC (2000) report, students should: (a) "give priority to *evidence*, which allows them to develop and evaluate explanations that address scientifically oriented questions" (p. 25), (b) "formulate *explanations* from evidence to address scientifically oriented questions" (p. 25), (c) formulate and revise scientific explanations and models using logic and evidence (p. 19), and (d) have a clear understanding that "scientific explanations emphasize evidence, have logically consistent arguments, and use scientific principles, models, and theories" (p. 20). Scientific inquiry, then, is fundamentally about acquiring relevant data, transforming that data first into evidence and then into explanations that can be conceived as answers to particular scientifically oriented questions (Duschl, 2003; Sandoval & Reiser, 2004).

It is assumed that constructing explanations helps students to understand the nature of scientific knowledge in terms of its connection to evidence, its uncertainty, and its subjectivity to change (Bell & Linn, 2000; Duschl, 2003; Sandoval, 2001, 2003; Sandoval & Reiser, 2004). Furthermore, it has been argued that having students write their scientific explanations helps them to reflect on what they are learning in a way that is not usual in oral exchanges, such as in classroom discussions (Tishman & Perkins, 1997). Writing may help students to think critically, and construct new knowledge by exploring the relationship between ideas and transforming rudimentary ideas into knowledge that is more coherent and structured (Bereiter & Scardamalia, 1987; Klein, 1999, 2004; Rivard & Straw, 2000). The construction of written scientific explanations should be considered, then, at the heart of scientific inquiry and should be emphasized in every science class in which scientific inquiry teaching is taking place.

In this paper, we analyze the quality of students' written explanations in eight classrooms, and explore the link between the quality of the explanations and the students' performance as measured in assessments focusing on the content of the science unit studied. More specifically, we asked the following questions: (a) How frequently did middle school students write explanations in response to the scientifically oriented questions they investigated in their science classrooms? (b) If they wrote explanations, what are the characteristics of these explanations? And (c) is there a link between the quality of students' written explanations and their performance in assessments, focusing on the content studied?

We also explore an instructional aspect by asking a fourth question: (d) What instructional prompts, if any, best promoted high quality explanations?

In what follows, we begin by providing the theoretical framework used to approach the analysis of the scientific explanations. We then report on how the data were collected and analyzed. Next, we describe characteristics of the explanations that students gave in their notebooks, and provide evidence of the link between quality of the students' scientific explanations and level of performance observed in different types of assessments. Finally, we focus on the prompts that teachers used to support students' construction of explanations.

Scientific Explanations and Student Learning

On Explanations

Explanations are answers to particular questions (Sandoval & Reiser, 2004). Explanations should connect patterns of data with claims about what the data mean. Three components are being cited frequently as essential in scientific explanations (Kenyon & Reiser, 2006; Kuhn & Reiser, 2004, 2006; McNeill & Krajick, 2006; Sandoval & Reiser, 2004; Tzou, 2006):

1. *Claim*: A testable statement or conclusion that answers a scientific question. A scientific claim typically focuses on what happened, or how or why something happened.
2. *Evidence*: Investigation data that helps to construct, support, and defend a claim. Originally, Toulmin (1958) named this component *data* to refer to the statements used as evidence to support the claim.
3. *Reasoning*: Statements given to justify claims. That is, they are justifications to show why the data count as evidence to support the claim through a conceptual and theoretical link. Toulmin (1958) used the term *warrants* instead of reasons.

Toulmin (1958) proposed three more components for an explanation, which he called argument[1]: (a) *qualifiers* or statements about how strong the claim is[2], (b) *backings* or the assumptions or reasons held, and (c) *rebuttals* or statements that contradict the data, warrants, qualifiers, or backings. In our analysis we have decided to follow the triad to simplify the recognition of explanations, and have considered qualifiers and rebuttals in a second level analysis (see following), but have ignored backings since often they are typically not made explicit in middle school science education (see Kelly, Drucker, & Chen, 1998; Simon, Erduran, & Osborne, 2006).

[1] Argument refers to the substance of claims, data, warrants, and backings that contribute to the content of an argument. Argumentation refers to the process of assembling these components (Simon, Erduran, & Osborne, 2006).

[2] That is, the conditions under which the claim holds true -- the universality of the claim.

What do we value in an explanation? An explanation should respond to the question that generated the inquiry. "The evaluation of the worth of any explanation is in relation to its value as an answer to the original question" (Sandoval & Reiser, 2004, p. 349). Second, claims should be warranted with evidence, and both should be clearly differentiated (Duschl, 2003, Kenyon & Reiser, 2006; Kuhn & Reiser, 2006; McNeill & Krajick, 2006; Sandoval, 2001, 2003). Third, evidence provided in an explanation should be valid, reliable, relevant, and sufficient to support the stated claim (Kenyon & Reiser, 2006; McNeill & Krajick, 2006; Sandoval, 2001, 2003). Fourth, an explanation should be coherent; that is, it should articulate causally claims for its inquiry questions (Sandoval, 2003). Finally, a sophisticated explanation involves a careful comparison between alternative explanations and logic chains to build up the claims using evidence.

The construction of explanations is influenced both by students' understanding of the science content and by their understanding of what constitutes a scientific explanation (McNeill & Krajcik, 2006; Sandoval, 2003). Therefore, it should be expected that explanations reflect students' level of understanding of the science content at hand.

Written Explanations and Student Learning

Writing in science to enhance student understanding of scientific content and processes has been supported by many researchers (Bass, Baxter, & Glaser, 2001; Baxter, Bass, & Glaser, 2000; Keys, Hand, Prain, & Collins, 1999; Rivard & Straw, 2000; Shepardson & Britsch, 1997). Indeed, the use of writing as a learning strategy has received considerable theoretical support (Applebee, 1984; Bereiter & Scardamalia, 1987). Improved learning as a result of writing has been attributed to the mental representations, strategies, and operations that take place while writing (Bereiter & Scardamalia, 1987; Klein, 1999). It has been proposed that students take a problem solving approach when writing. They set goals and then follow strategies to construct and transform discursive knowledge to match their perceived requirements of the writing task (Bereiter & Scardamalia, 1987).

Brown and Campione (1990) argued that asking students to write explanations push them to evaluate, integrate and elaborate knowledge in new ways that positively impacts their learning. Furthermore, being involved in the process of explaining to themselves or to others helps develop competence (Chi, 2000). Certainly, asking students to write their explanations has proven to have a positive impact on their science learning and transfer of knowledge (Boscolo & Mason, 2001; Keys, 2000; Klein, 2004). The rationale is that having students present their explanations in written language engages them in a specific type of reflection that is not natural in oral exchanges (Tishman & Perkins, 1997). While oral discourse is

divergent, highly flexible, and requires little effort from students, written discourse is convergent, more focused, and places greater cognitive demands on the writers (Rivard & Straw, 2000). Although writing and talking are complimentary modalities, the use of writing as an instrument for constructing explanations underlies the personal, rather than social, construction of knowledge. It has been argued that writing in science helps to structure, categorize, and acquire the disciplined characteristics of scientific knowledge (Halliday & Martin, 1993).

Characteristics of written explanations. What is the most appropriate way to communicate a scientific explanation in writing? Is there a functional convention defined by certain parameters? The research community is divided (Klein, 2006; Prain, 2006). Some consider there to be grammatical resources that have been used to construct, represent, and disseminate the knowledge of science (Halliday & Martin, 1993; Martin, 1993a). Therefore, various genres have been developed to provide structures to represent scientific reasoning, argument, and discourse. These genres are viewed as representing the epistemic essence of science as a discipline and field of study (Prain, 2006).[3] The genres are adapted to the different aspects of science that are being addressed. For example, explanations are considered a genre in which there is a high percentage of action timeless verbs and the actions are organized in a logical sequence (Martin, 1993a). They give rise to technical terms and consequential relationships (Martin, 1993b).

For others, students should be encouraged to write in diverse forms as long as they can relate emerging knowledge and technical vocabulary with their previous knowledge and experiences. Under this perspective, students should use and engage their linguistic resources to develop and demonstrate understanding without considering specific parameters such as those suggested by the scientific genres. Therefore, they advocate for expanding the purposes and types of writing in science beyond the traditional science genres (e.g., Bereiter & Scardamalia, 1987; Klein, 1999; Prain, 2006). For example, asking students to write their initial ideas or conceptions about a phenomenon, write an interpretation of what appears to be an interesting event, write to communicate what has not been understood, or write to make comments, are all considered appropriate ways to encourage students to write in science (Boscolo & Mason, 2001). Science literacy education should involve "…whatever activities with whatever media [that] may help students to become scientifically literate" (Klein, 2006, p. 146).

[3] A genre represents the socially accepted parameters for appropriate content generation, organization, stylistic choices, and voice (Rijlaarsdam, Couzijn, Janssen, Braaksma, & Kieft, 2006).

We believe that students' explanations should have some critical characteristics that show the main purposes for what explanations are constructed. Therefore, we argue that an explanation should start with its relation with the problem, research question, purpose, goal, or the hypothesis investigated. The rationale is that the value of an explanation is in relation to how it answers the original question (Sandoval & Reiser, 2004). We also think that students' explanations should be clear, complete, and precise so they can express in a comprehensible way the reasoning connecting the evidence and the claim. Finally, we argue that explanations should contain the suitable technical terms, not only because they optimize communication (Martin, 1993a), but also because technical terms reflect somehow how well the students understand the core concepts at hand (e.g., the use of mass instead of weight). In our analysis we have focused on these three features of written communication. In this paper we do not focus on the communication characteristics of the explanations, but all the aspects discussed were considered in the scoring approach.

Explanations and Instructional Strategies

Research in cognitive psychology has shown that different writing tasks invoke different cognitive strategies for students to process and retain information (Bereiter & Scardamalia, 1987; Klein, 2004). The issue, then, is to identify types of instructional prompts that can provide students with support for constructing explanations that have an impact on their learning (Hand & Prain, 2006). Different prompts for this purpose have been studied: writing heuristics (Keys, 2000), explanation guides (Sandoval & Reiser, 2004), explanation frameworks (Kenyon & Reiser, 2006), integration frameworks (Bell & Linn, 2000), and written scaffolds (McNeill & Krajcik, 2006). Some are generic (e.g., an explanation skeleton; Kenyon & Reiser, 2006; Keys, 2000; McNeill & Krajcik, 2006), and others, content-based (e.g., questions specific to the investigation conducted; Bell & Linn, 2000; McNeill & Krajcik, 2006; Sandoval & Reiser, 2004). Some help students to focus on the possible content of explanations (e.g., Sandoval & Reiser, 2004), whereas others help students focus on claims and evidence (e.g., Kenyon & Reiser, 2006). In some cases, the prompts are computer-based (Bell & Linn, 2000; Sandoval & Reiser, 2004); others engage teachers in the acquisition or acquaintance of new forms of classroom discourse, implementation of prompts, or implementation of improved curricula with the prompts (e.g., Kenyon & Reiser, 2006; McNeill & Krajcik, 2006).

All these studies have involved the implementation of a particular treatment and the evaluation of its impact in promoting students' construction of explanations. The treatment can be conceived as the implementation of an instructional strategy (e.g., the type of prompt or a curriculum) to help students to construct explanations, assuming that such construction

will affect students' learning as well as their views of the nature of science (Bell & Linn, 2000; Keys, 2000).

The study described in this paper differs from the others in that it focuses on whether the construction of explanations naturally occurs in science inquiry classrooms without implementing any treatment, and if so, what impact this construction may have on student learning. More specifically, we study the quality of the explanations written by the students and explore whether a link between the quality of the explanations, and the students' learning and achievement could be established. We also explore the types of prompts used naturally by the teachers to engage students in the construction of explanations.

The Context of the Study

The data analyzed in this study is part of a larger project involving a collaboration between the Stanford Education Assessment Laboratory (SEAL) and the Curriculum Research and Development Group (CRDG) at the University of Hawaii at Manoa. The project was conducted using the Foundational Approaches in Science Teaching (FAST) middle-school science curriculum developed by the CRDG (Pottenger & Young, 1992). The project focused on the effects of formal embedded assessments on students' learning on *relative density* (Shavelson & Young, 2000).[4] All teachers participating in the project were asked to provide their students' science notebooks, among other artifacts, at the end of the school year.

We used students' science notebooks as the main source of information to analyze the characteristics of the explanations. The rationale is that notebooks are generated during the process of instruction and somehow reveal what students do in their classrooms. Some studies have indicated that students' notebooks reflect with great fidelity what students do and what teachers focus on in the science class (Alonzo, 2001; Baxter, Bass, & Glaser, 2000).

We maintain, then, that science notebooks reflect, at least partially, the instructional tasks carried out in a science class. Furthermore, they offer a window into students' thinking and learning, and they can provide some evidence of teachers' communications with students about their progress (Ruiz-Primo, Li, Ayala, & Shavelson, 1999, 2004; Ruiz-Primo & Li, 2004; Ruiz-Primo, Li, & Shavelson, 2001; Warren-Little, Gearhart, Curry, & Kafka, 2003).

In previous studies it has been reported that students' explanations can be found in their notebooks, some in the form of conclusions of their investigations, some others as responses

[4] For more information on this project please see Shavelson & Young, 2000.

to teacher's questions, and still others just as explanations (Aschbacher & Alonzo, 2006; Ruiz-Primo, Li, Ayala, & Shavelson, 1999, 2004; Ruiz-Primo & Li, 2004; Ruiz-Primo, Li, & Shavelson, 2001). In what follows we describe in more detail the approach we used in characterizing students' explanations.

Approach to Analyzing Scientific Explanations

We focus on students' explanations from two perspectives: the characteristics of the individual explanations and the characteristics of the instructional devices used at the classroom level to construct the explanations. At the individual level, we examine the explanations in terms of: (a) their quality, and (b) the level of students' understanding demonstrated in the explanations.

Students' explanations at the classroom level capture the type of prompts used to engage students in the construction of explanations, for example, whether teachers provided questions or skeletons to guide students in the construction of their explanations.

Quality of Explanations

We evaluated the quality of student explanation at two levels. The first level focused on the three components of the explanation defined in the previous section (i.e., claim, evidence, and reasoning). The second level focused on other characteristics that could be expected in an explanation that is complete and exceptional.

Quality of the components of the explanation. This level focuses on the *function* of an explanation using its three components as the basis: claim, evidence, and reasoning. We first identified whether the explanation found in the notebook had the three components. For each explanation component identified we evaluated its quality by addressing a set of questions focusing on different aspects (Table 1).

To evaluate the *quality of the claim*, we focused on whether the claim corresponded to the main issues tapped by the investigation at hand. We called this aspect *focus*. We also captured the accuracy of the claim, that is, if the statements were scientifically sound. For example, a claim could address the two main issues expected according to the research question, but one of them was correct and the other one was not (i.e., all main issues in the claim addressed, but they are partially incorrect).

We coded three aspects of the *quality of the evidence*: *type* (i.e., What type of evidence did the student provide, anecdotal, concrete examples, or investigation-based?), *nature* (i.e., Did the student focus on patterns of data or isolated examples?), and *sufficiency* (i.e., Did the student provide enough evidence to support the claim?). It has been argued that the evidence

that students select to include in their explanations likely reflects their ideas about what counts as important to understand the phenomenon at hand, what is relevant to include in their explanations, and what data they actually understand (Sandoval, 2001).

We focused on two aspects of the *quality of the reasoning*: *alignment* (i.e., Is the evidence related to the claim?) and the *type of link* (i.e., How was the provided evidence connected to the claim?).

Table 1

Examples of Coding Questions and Criteria to Score Quality of Explanations

Component	Examples of coding questions	Examples of scoring criteria
Claim	How does the claim accurately address the main question or issue tapped in the investigation conducted?	• Does not address • Partially addresses • Accurately addresses all main ideas
Evidence	What type of evidence did the student provide?	• No evidence • Anecdotal/opinion/everyday examples • Investigation data
	What form of evidence did the student provide?	• Qualitative data pattern • Quantitative data pattern • Specific examples
	If specific examples, how many were provided?	• One data point • Two data points • More than two data points
Reasoning	Was the evidence provided aligned with the claim?	• No • Partially • Completely
	How was the provided evidence connected to the claim?	• No link • Link indicated by connected words only • Elaborated link

Extended quality of the explanations. This dimension focused on whether students gave consideration to other characteristics of an explanation besides its basic components. We evaluated whether students addressed: (a) issues related to the quality of the evidence or the reasoning (e.g., Is there any evidence that the student evaluated the quality of the data collected? If so, what type of evaluation was provided?), (b) alternative explanations in response to the question at hand (i.e., Did the student consider counter-arguments or alternative explanations when building the claims or reasoning through?), and (c)

implications of the explanation (e.g., Did the student discuss some applications of what was learned?). Students' discussion on each of these aspects was coded for its relevance (Table 2).

Table 2

Examples of Coding Questions and Criteria to Score Extended Quality of Explanations

Extended component	Examples of coding questions	Examples of scoring criteria
Quality of evidence and reasoning	Did the student evaluate the quality of the evidence? If yes, what type of evaluation was provided?	• Measurement error • Strength/weakness of the link
	How relevant and appropriate was the discussion?	• Irrelevant • Relevant but superficial • Relevant and complete
	Did the student consider alternative explanations?	• No • Yes, but inappropriate • Yes, and appropriate
Implications	Did the student consider some implications of what it was learned in the investigation?	• Connection with other topics • Connection with other topics studied • Practical application of the findings • Limitation of the findings

Student's Level of Understanding

To identify students' level of conceptual understating demonstrated in their written explanations, we used a conceptual progress trajectory of density (SEAL, 2003) as reflected in the FAST science unit implemented in the larger study (Table 3). This trajectory could be used across different investigations of the unit to determine whether the students were achieving the expected level of understanding at the different stages of the unit.

Table 3

Coding Framework for Students' Developing Understanding of Sinking and Floating (Adapted from Yin, 2005)

Level	Level of understanding	Description: *Students' explanation focused on …*
6	Relative density	The comparison of the density of *both* the object and liquid
5	Density	The density of *either* the object *or* the liquid
4	Mass *and* Volume	Mass and volume *together* influencing sinking and floating
3	Mass *or* Volume	*Either* mass *or* volume influencing sinking and floating
2	Naïve science conception	Chemical/material/component, force/pressure/buoyancy
1	Alternative conception	Air, medium size, shape, hole, hollow/solid

Instructional Prompts

We also captured how the constructed explanations were prompted. For example, we asked: Was the explanation part of the conclusion section of the investigation report? Or, was the explanation provided as a response to a teacher question or set of questions? We created nine categories to classify the type of prompt found (e.g., teacher questions). With this information we wanted to track the characteristics of prompts at the classroom level. Were they guided and content specific, or were they generic? The purpose of this analysis was to define whether some prompts tended to be more effective that others in supporting students in constructing explanations.

Method

Participants

Seventy two middle school students and eight science teachers in eight different schools, in five states from the larger project participated in the study. Table 4 provides general information about the teachers and classrooms in the study.

Table 4

Classrooms General Characteristics

Characteristic	Classrooms							
	1	2	3	4	5	6	7	8
Teacher								
Highest degree	ME	BA	ME	BA	BS	BS	MS	MA
Teaching experience (years)	2	18	23	3	6	14	12	22
Teaching science (years)	2	17	10	1	3	14	12	6
Teaching FAST (years)	2	12	1	1	3	7	12	2
	Number of students							
Class Size	29	22	20	25	21	27	29	24
	Percentages							
School characteristics								
Ethnicity								
African American	2	1	1	1	0	9	4	23
American Indian	12	0	1	0	0	11	2	0
Asian	4	91	3	0	1	2	20	7
Hispanic	4	1	3	1	0	3	7	1
White	79	7	92	98	98	75	67	68
Mathematics proficiency level[a]	34	30	55	77	32	80	24	39

[a]This information was captured from each school district's web site at the time when the data for the larger project was collected. It refers to the percentage of students who met the state mathematics standards.

Curriculum

All teachers implemented FAST. The first unit of FAST, *Properties of Matter,* supports students in the development of *relative-density* based explanations of sinking and floating through 12 inter-related investigations.

This paper focuses on Investigation 7, *Floating and Sinking Objects*. In this investigation students measure the mass of sinking and floating objects, and their displaced volume after being placed in water. The main goal of this investigation is to determine the relationship between the mass of an object, the volume of displaced water, and sinking/floating (Pottenger & Young, 1992). We focused on this investigation because it is critical in the development of understanding the concept of density. In this investigation students, for first time, studied the factors of mass and volume on sinking and floating after focusing on each of the factors separately. We reasoned that this investigation could have the

biggest impact on the students' understanding for the next investigations, and therefore, on their performance at the end of the unit.

Investigation 7 focused on whether or not it was possible to predict the displaced water of floating and sinking objects if the mass of the objects was known. In this investigation students, working in small groups, massed floating and sinking objects, measured the volume of displaced water, recorded the data collected in tables within the small group and across groups in the class, graphed the class data for both factors (mass and displaced volume), and determined the relation between mass and volume of floating and sinking objects.

After conducting Investigation 7, it is expected that students make the leap of considering only mass (e.g., more mass more sinking - Investigation 4) or only volume (less volume less sinking - Investigations 5 and 6) as interrelated factors determining sinking or floating (e.g., an object with more mass than volume will sink). Students' explanations, then, should show conceptual understanding at Level 4 of the trajectory (see Table 3).

It is important to mention that the explanations expected in this investigation do not focus on causal mechanisms (chains of a cause and consequence effects), but more on the articulation of the students' understanding about an occurred event (Kuhn & Reiser, 2004). This means that the claims were expected to focus on describing what happened (i.e., for floating objects it is possible to predict the amount of displaced water, but not for sinking objects) more than on identifying a critical factor in a causal relationship.

Sources of Information

As mentioned, teachers were asked to provide their students' science notebooks at the end of the school year. The science notebooks included reports of the investigations carried out by the students. Each student notebook was analyzed on six aspects of the reported FAST investigations: Problem, Vocabulary, Background, Method, Reporting Results, and Conclusions. This paper focuses on the analysis of the "Conclusions" of Investigation 7 reported in the students' science notebooks, which included explanations embedded in the conclusion section of the reports or in any other entry of the notebook that presented a written explanation (e.g., answers to questions provided by the teachers).

Within each classroom, nine students' notebooks were randomly selected from strata based on the students' scores on the multiple-choice test administered at the end of the unit (three high-, three medium-, and three low-proficient).

To explore the link between the quality of students' explanations and their performance at the end of the unit, the students were administered four types of assessments in a pre-

test/post-test design as part of the larger project (see Yin, 2005 for details). In the pre-test, students were administered a 36 multiple-choice test. In the post-test, students were administered the multiple-choice test, a predict-observe-explain assessment (POE), a performance assessment (PA), and an open-ended question. The multiple-choice test included almost all of the instructional objectives covered in the 12 investigations, including items focusing on issues explored in Investigation 7. The internal consistency of the multiple-choice test was .858 (see Yin, 2005).

The POE had three parts: (a) observe an experimental setting based upon concepts already learned, (b) make a prediction, and (c) reconcile the prediction with the actual outcome of the experiment. The assessment was developed around the density of a bar of soap in the context of sinking and floating (Yin, 2005). Students first observed that the whole soap sank in a container of water and then they were asked to predict whether a piece of that soap (about a fourth) would sink or float. The assessor then put the piece of the soap in water and students were asked to reconcile their prediction with the actual outcome of the experiment. Interrater agreement for this assessment was 92.2 (see Yin, 2005)

In the PA, students were supplied with equipment (four blocks with different densities, water, graduated cylinder, rulers, overflow can, and other necessary supplies) and were asked to: (a) find the density of a block with a given mass; and (b) find the density range of a mystery liquid. A set of person (p) by rater (r) G studies indicated that the PA could be reliably scored (averaged ρ^2 coefficient was .825; see Yin, 2005).

The short open-ended prompt asked students a central question across all the 12 investigations studied: Why do things sink and float? We named this assessment, the WTSF assessment. Interrater agreement for this assessment was 87.2 (see Yin, 2005)

Coding System

Students' science notebooks were analyzed with the assistance of an *Access* computer program. The scoring approach is based on the idea of "*hierarchical trees*." That is, aspects to be scored within each notebook element are hierarchically organized; some aspects are subordinated to others. If an aspect at a higher level is not found, the aspects that are subordinated to it are assumed not to be there either. For example, the computer program first asks: Is there any evidence that there is an explanation in the notebook at hand? If the response is yes, a set of questions is asked regarding the quality of the explanation: components of the explanation, quality of the claim, quality of the evidence provided, and quality of the reasoning and the like (Figure 1). If the response is no, the scorer skips this aspect and continues to score the next aspect of the notebook.

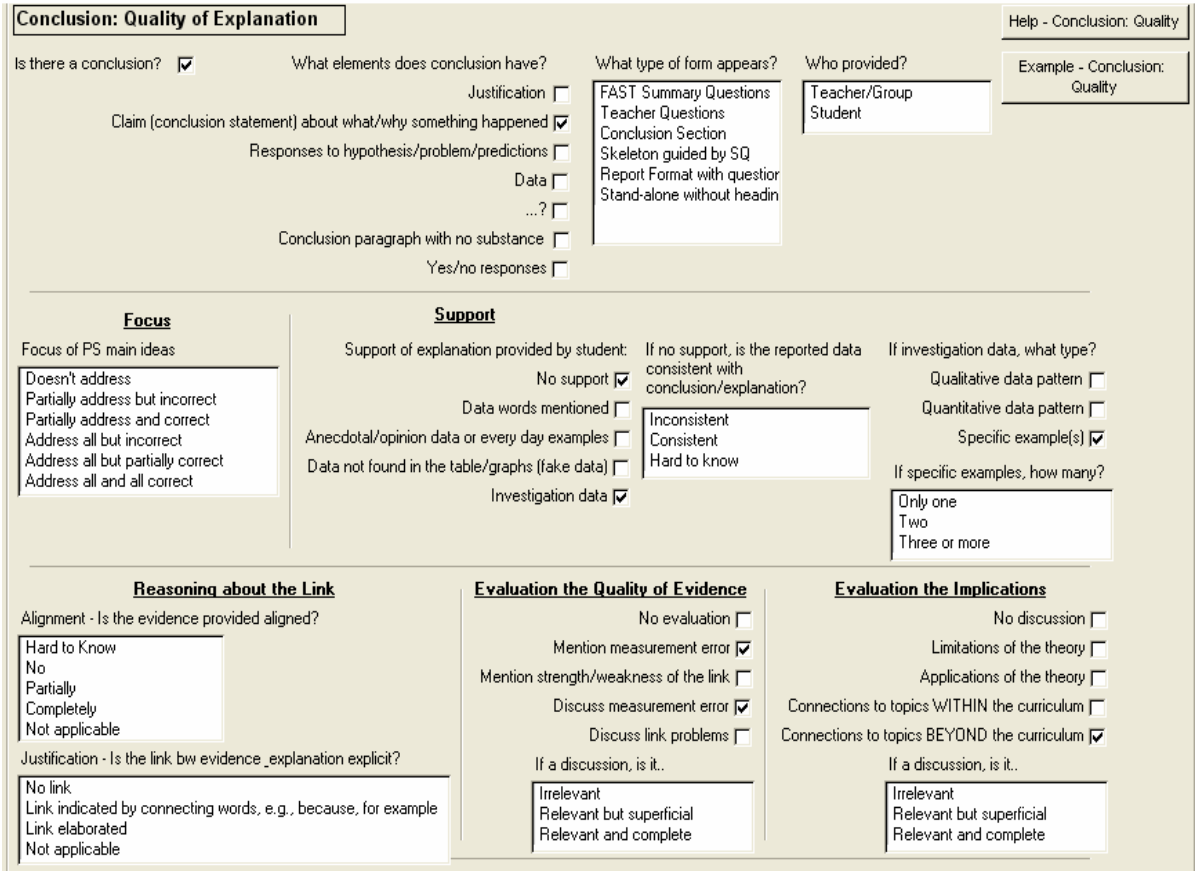


Figure 1. Microsoft Access screen with most of the options for scoring quality of explanations found in the science notebook entries.

Figure 2 provides an example of a notebook entry that reports Investigation 7, and Figure 3 shows the different Access screens that appeared as the scoring process proceeded. Figure 2 provides the complete entry to show how some parts of the entry (e.g., the data table) are used to respond to some questions related to the explanation. We assume this example aids in explaining the scoring procedure.

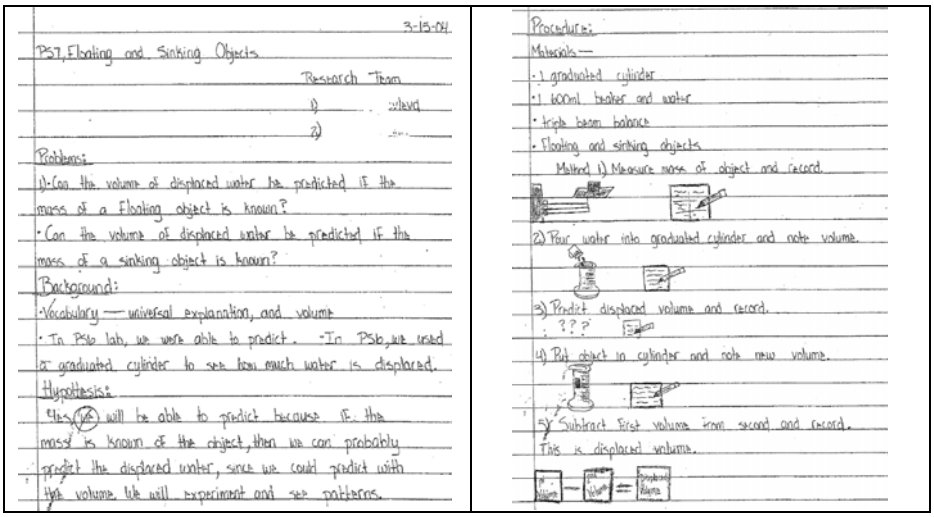
As mentioned, the first question asked is whether there is evidence that a component is present (e.g., problem). If the box is checked, a set of questions about the quality of that aspect is asked. The approach continues asking questions about the different aspects until the Conclusion aspect is reached (see Figure 2e and f): "Is there a conclusion?" (see Figure 3a).

When the box was checked to respond to the question (i.e., yes, there is a conclusion), a list of boxes dealing with the explanation components appears on the screen as well as two selection menus (i.e., What type form? and Who provided it?; see Figure 3b). Three selections were made in this step: (a) The data box was checked (i.e., "The data shows for the floaters, the mass of the object is the same as displaced volume. For the sinkers the volume

of displaced water is less than the mass of the object. For example, a floater with…"; see Figure 3b), (b) "Conclusion" section was selected, and (c) "Student provided" was selected.[5]

If the "Data" box is selected another set of boxes appears on the screen looking at the "Type of Evidence." Two boxes were checked, Data not found in table/graph (i.e., "For example floater with 13g displaces 13 mL, also a floater with 5g displaces 5mL.") and "Investigation data" (i.e., "Another example is a sinker with 6g displaces 5mL" which it was found on the graph, but not in the table). If "Investigation data" is selected, a set of boxes appears dealing with the nature of the data (see Figure 3c). "Qualitative pattern" and "Examples" were selected. Once "Examples" is selected, another menu appears (i.e., Only one, Two, or Three or more examples). Three or more examples was selected in this menu. The procedure continues until every element in the scoring system dealing with "Conclusion" is considered (see Figure 1).

[5] Rules were developed to define who provided a component at hand. If, in a sample of four notebooks within the same classroom, the information was the same across students, then "Teacher provided" was selected.



3-15-04

PS7, Floating and Sinking Objects

Research Team

1) ...steve

2) ...

Problems:

1) Can the volume of displaced water be predicted if the mass of a floating object is known?

• Can the volume of displaced water be predicted if the mass of a sinking object is known?

Background:

• Vocabulary — universal explanation, and volume

• In PS6 lab, we were able to predict. • In PS6, we used a graduated cylinder to see how much water is displaced.

Hypothesis:

We (we) will be able to predict because if the mass is known of the object, then we can probably predict the displaced water, since we could predict with the volume. We will experiment and see patterns.

(a)

Procedure:

Materials —

• 1 graduated cylinder

• 1 600ml beaker and water

• triple beam balance

• floating and sinking objects

Method 1) Measure mass of object and record.

2) Pour water into graduated cylinder and note volume.

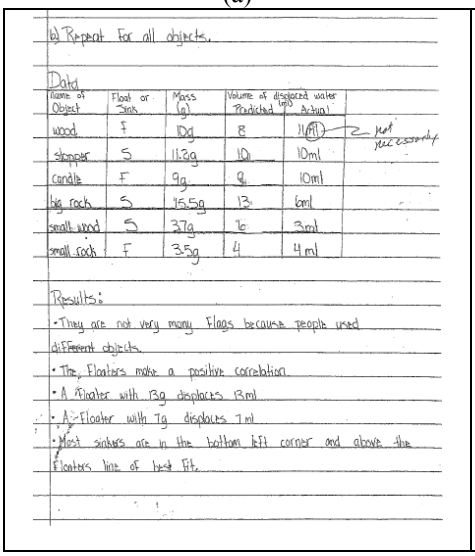
3) Predict displaced volume and record.

??? 5ml

4) Put object in cylinder and note new volume.

5) Subtract first volume from second and record. This is displaced volume.

(b)



6) Repeat for all objects.

Data

Name of Object	Float or Sink	Mass (g)	Volume of displaced water Predicted (ml)	Volume of displaced water Actual (ml)
wood	F	10g	8	11ml
stopper	S	11.3g	10	10ml
candle	F	9g	8	10ml
big rock	S	15.5g	13	6ml
small wood	S	37g	16	3ml
small rock	F	35g	4	4 ml

not necessarily

Results:

• They are not very many flags because people used different objects.

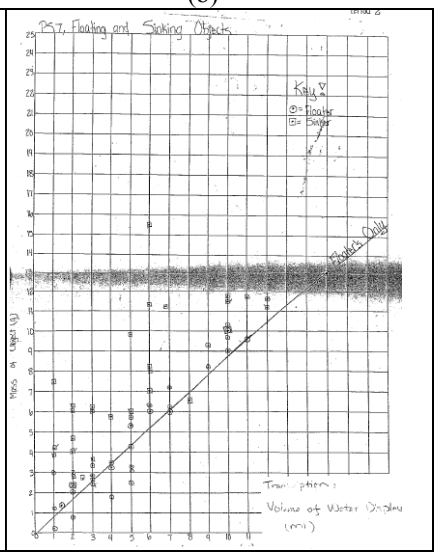
• The floaters make a positive correlation

• A floater with 13g displaces 13ml

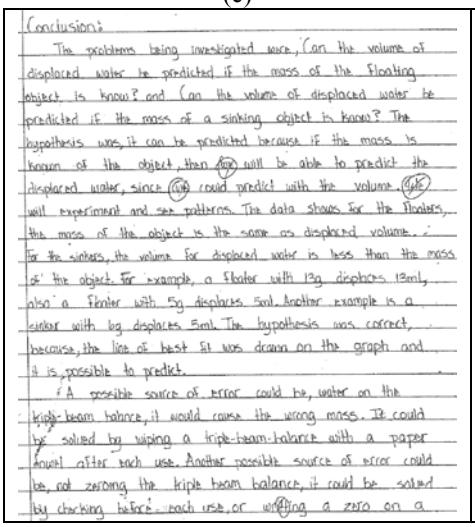
• A floater with 7g displaces 7 ml

• Most sinkers are in the bottom left corner and above the floaters line of best fit.

(c)



(d)

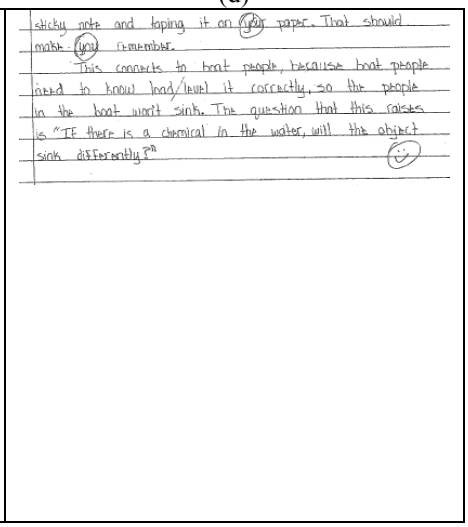


Conclusion:

The problems being investigated were, Can the volume of displaced water be predicted if the mass of the floating object is know? and Can the volume of displaced water be predicted if the mass of a sinking object is known? The hypothesis was, it can be predicted because if the mass is known of the object, then (we) will be able to predict the displaced water, since (we) could predict with the volume. (We) will experiment and see patterns. The data shows for the floaters, the mass of the object is the same as displaced volume. For the sinkers, the volume for displaced water is less than the mass of the object. For example, a floater with 13g displaces 13ml, also a floater with 5g displaces 5ml. Another example is a sinker with 6g displaces 5ml. The hypothesis was correct, because, the line of best fit was drawn on the graph and it is possible to predict.

A possible source of error could be, water on the triple beam balance, it would cause the wrong mass. It could be solved by wiping a triple-beam-balance with a paper towel after each use. Another possible source of error could be, not zeroing the triple beam balance, it could be solved by checking before each use, or writing a zero on a

(e)



sticky note and taping it on (my) paper. That should make (me) remember.

This connects to boat people, because boat people need to know load level it correctly, so the people in the boat won't sink. The question that this raises is "If there is a chemical in the water, will the object sink differently?"

(f)

Figure 2. Example of a student's notebook entry of Investigation 7.

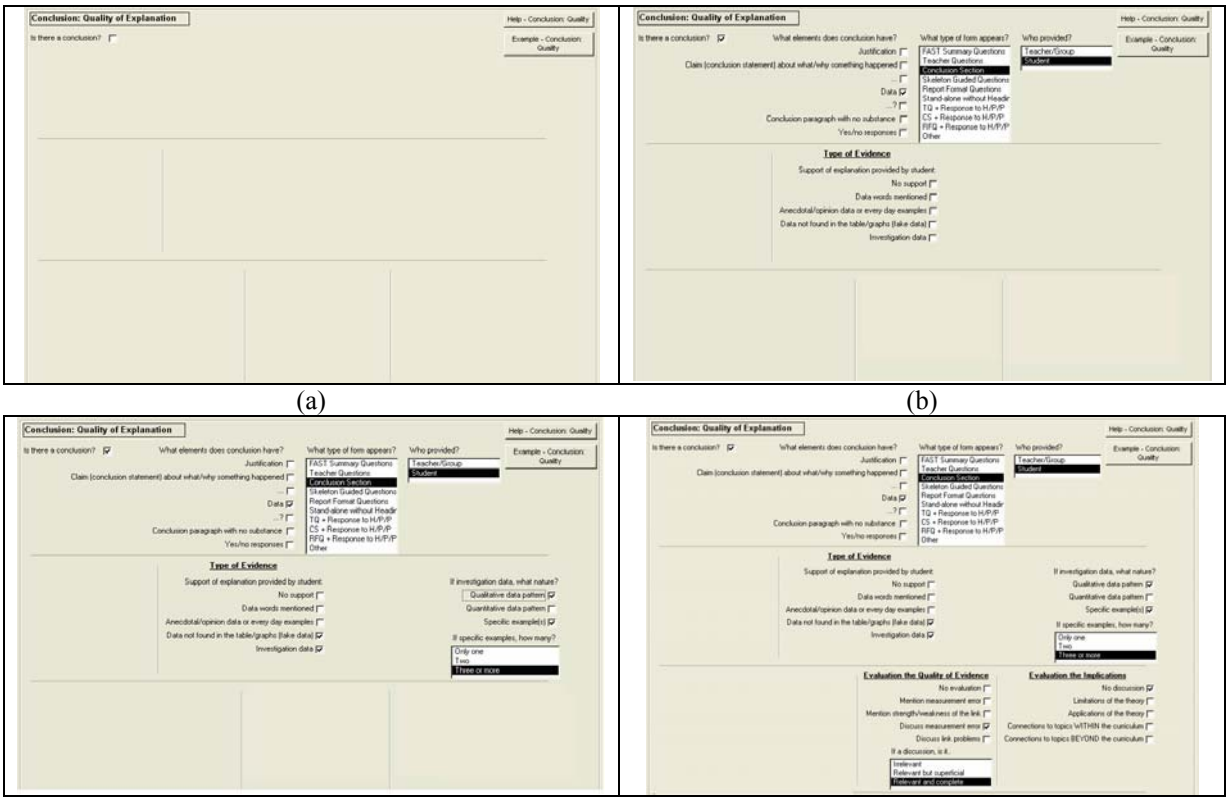


Figure 3. A sample of the Microsoft Access screens with the options selected for scoring the student's notebook conclusion entry provided in Figure 2.

Agreement and Inter-Rater Reliability

Each entry for Investigation 7 was coded and scored using the approach described. To assess the consistency among the raters, 12 notebooks selected randomly from the 8 classrooms were scored by three raters. We developed six types of scores to account for the quality of the explanations: type of explanation, focus of the claim, quality of evidence, alignment between claim and evidence, level of understanding, and a composite score. To examine the generalizability of the scores across raters, six person (p) x rater (r) G studies were carried out, one for each type of score (Table 5). Results indicated that with carefully defined scoring rules and well trained raters, students' explanations from science notebooks can be reliably scored.

Table 5

Relative and Absolute G Coefficients with 3 Raters

Source of variability	Type of explanation		Focus of the claim		Quality of evidence		Alignment		Level of understanding[a]		Composite score	
	EVC	%	EVC	%	EVC	%	EVC	%	EVC	%	EVC	%
person (p)	0.86	94	2.33	94	1.32	99	0.28	86	0.64	78	5.96	97
rater (r)	0.00	0	0.02	0	0.01	1	0.00	0	0.02	3	0.00	0
pr, e	0.05	6	0.14	6	0.00	0	0.04	14	0.15	19	0.20	3
$n_r = 3$												
ρ^2	0.94		.94		1.00		0.86		0.81		0.97	
φ	0.94		0.94		0.99		0.86		0.78		0.97	

[a]Only two raters were selected for estimating the variance components for this type of scores.

Results

In this paper we asked the following questions: (a) How frequently did middle school students write explanations in response to the scientifically oriented questions they investigated in their science classrooms? (b) If they wrote explanations, what are the characteristics that these explanations have? And (c) is there a link between quality of students' written explanations and their performance in assessments focusing on the content studied? We also explore an instructional aspect about explanations and asked a fourth question: (d) What instructional prompts, if any, best promoted high quality explanations? In what follows we respond to each of the question posed.

Writing Explanations Practices across the Eight Classrooms

How frequently did middle school students write explanations in response to the scientifically oriented questions they investigated in their science classrooms? To respond to this question we considered four forms of explanations that students presented: (a) complete explanations (the triad identified), (b) only claim and evidence, (c) only claim, and (d) only data.

From the sample scored across the eight classrooms (n = 72) we found that only 18.1% of students provided explanations with the three expected components (i.e., claim, evidence, and reasoning). Only 12.5% provided claims with supporting evidence. The majority (40.3%) provided only claims without any supporting evidence, and 9.7% provided only data. Also, 19% of students did not provide any "form" of explanation. These results suggest that claim is the easiest component for students to construct and for teachers to focus on.

Complete explanations with the three components discussed, were identified mainly in two classrooms, 5 and 7 (Table 6). Classroom 2 showed the highest percentage of students who provided only claim and evidence (77.8%), and Classroom 3, the highest percentage of only evidence (66.7%). Notice that Classroom 6 is the classroom with the highest percentage of students' notebooks that did not provide any indication of any type of "explanations." Teachers across classrooms focused on different aspects of explanations.

Table 6

Percent of Students' Explanations Provided by Type and by Classroom

	Classrooms							
Type of "explanation"	1	2	3	4	5	6	7	8
Explanations	11.1	0.0	0.0	0.0	55.6	0.0	77.8	0.0
Claims & evidence	0.0	77.8	11.1	0.0	11.1	0.0	0.0	0.0
Only claims	66.7	22.2	22.2	77.8	0.0	33.3	11.1	88.9
Only evidence	0.0	0.0	66.7	0.0	0.0	0.0	11.1	0.0
No explanation	22.2	0.0	0.0	22.2	33.3	66.7	0.0	11.1

Quality of the Components of the Explanation

In this section we focus on those students who provided at least one component of explanations. First, we focus on analyzing the quality of the explanation components, and then on the analysis of the extended quality of the explanations.

Claims. To analyze the quality of the claims we focused on looking at three major elements of the students' explanations according to the focus of Investigation 7: (a) relating mass of an object to the volume of displaced water, (b) qualifying the relationship for floaters, and (c) qualifying the relationship for sinkers. For each element, we also considered their correctness (e.g., for floating object, if the mass is known, it is possible to predict the amount of water they will displace). The different categories and the percentages observed across classrooms are presented in Table 7.

Results indicate that the highest percentages of the claims which focused correctly on the expected elements were found in Classrooms 5 and 7 followed by Classroom 4. However, it is important to note that a high percentage of the claims did not address any of the expected elements (e.g., "more mass, more depth of sinking," a correct claim, but failed to address this

investigation). What was, then, the students' understanding of the investigation conducted? What was the purpose of it?

Table 7

Quality of Students' Claims by Characteristic and Classroom in Percentages

Focus of claims (elements considered)	Classrooms							
	1	2	3	4	5	6	7	8
All correctly addressed	0.0	11.1	11.1	33.3	55.6	0.0	77.8	0.0
All but partially correct	11.1	55.6	0.0	0.0	11.1	0.0	0.0	11.1
All incorrectly addressed	0.0	33.3	0.0	11.1	0.0	0.0	0.0	11.1
Some correctly addressed	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Some incorrectly addressed	0.0	0.0	0.0	11.1	0.0	0.0	0.0	0.0
Did not address any element	66.7	0.0	22.2	22.2	0.0	33.3	11.1	66.7
No claims only evidence	0.0	0.0	66.7	0.0	0.0	0.0	11.1	0.0
No explanation	22.2	0.0	0.0	22.2	33.3	66.7	0.0	11.1

Evidence. Although evidence can take many forms (e.g., background information; Sandoval, 2003), due to the nature of the investigation at hand, evidence refers here to numerical data. Students could provide evidence as patterns observed (e.g., "for floating objects, the relationship between the mass of the object and the volume of water displaced is of 1 to 1"), or one or more examples (e.g., "the cork, a floating object, massed 6 g and it displaced 6 mL of water"). In this section we consider those students who provided data, regardless of whether or not the data were used as evidence to support their claims.

We focused on three aspects of the quality of the evidence provided: type, nature, and sufficiency of data. Table 8 provides information of the three aspects considered and the categories used within each of them to analyze the evidence. Results indicate that most of the students did not provide any evidence (59.7% across the eight classrooms). From those students who provided evidence, 22.2% provided data collected during the investigation; 8.3% combined data collected in the investigation with artificial data; that is, data that could not be found either in the class tables or class graphs. We did not find any case in which evidence was based only on artificial data or anecdotal data.

Table 8

Characteristics of the Evidence Provided by Aspect of Quality and Classroom in Percentages

Aspects of quality of evidence	Classrooms							
	1	2	3	4	5	6	7	8
Type of evidence provided								
Investigation data	11.1	0.0	55.6	0.0	66.7	0.0	44.4	0.0
Investigation & artificial data	0.0	0.0	22.2	0.0	0.0	0.0	44.4	0.0
Artificial data	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Anecdotal data	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Word "data" only mentioned [a]	0.0	77.8	0.0	0.0	0.0	0.0	0.0	0.0
No evidence	88.9	22.2	22.2	100.0	33.3	100.0	11.1	100.0
Nature of evidence provided								
Qualitative pattern only	11.1	0.0	0.0	0.0	0.0	0.0	11.1	0.0
Quantitative pattern only	0.0	0.0	22.2	0.0	33.3	0.0	11.1	0.0
Qualitative pattern & examples	0.0	0.0	0.0	0.0	11.1	0.0	22.2	0.0
Quantitative pattern & examples	0.0	0.0	55.6	0.0	22.2	0.0	11.1	0.0
Data examples only	0.0	0.0	0.0	0.0	0.0	0.0	33.3	0.0
Word "data" only mentioned [a]	0.0	77.8	0.0	0.0	0.0	0.0	0.0	0.0
No evidence	88.9	22.2	22.2	100.0	33.3	100.0	11.1	100.0
Sufficiency of evidence								
Sufficient	11.1	0.0	77.8	0.0	66.7	0.0	55.6	0.0
Insufficient	0.0	0.0	0.0	0.0	0.0	0.0	33.3	0.0
Irrelevant	0.0	77.8	0.0	0.0	0.0	0.0	0.0	0.0
No evidence	88.9	22.2	22.2	100.0	33.3	100.0	11.1	100.0

[a]When students just mentioned data, table, or graph in their explanations instead of presenting or describing data, we coded as Word "data" only mentioned.

Classroom 3, 5, and 7 were the classrooms that show the highest percentage of students who provided evidence based on information they collected during the investigation. Classroom 2 was the single classroom in which we observed that the only support for the claims presented was the word data (e.g., "that is what the data say").

It is important to notice that in three of the classrooms (4, 6, and 8) we did not find any student who provided any type of evidence, but only claims. It is hard to know whether students were encouraged to provide any evidence at all.

Notably, 26% of the students focused on *patterns of data*; some of these students provided not only information about the pattern observed, but also data that exemplified the pattern. It seems that in Classrooms 3, 5, and 7, students were encouraged to focus on patterns of data, although not always to think about this pattern as evidence that could support their claims (Classroom 3). Only in Classroom 7 did we find students who provided only data examples as evidence.

When examples were provided, either accompanying a pattern or not, the majority of these students (75%) provided only one example, and only few (25%), provided two. None of the students in the sample analyzed provided three or more examples.

We judged sufficiency based on the nature of the data. When investigation-based patterns were provided (regardless of whether accompanied by examples), we considered that the data was sufficient and relevant to be used as evidence to support claims. When only examples were provided, if they were appropriate and more than three in number, we considered the evidence insufficient but relevant. When the "word" data was the only piece of information provided, we considered it irrelevant and insufficient for supporting the claims. Classrooms 3, 5, and 7 showed the highest percentage of sufficient evidence to support claims, whether or not these claims were stated (see Table 5, Classroom 3). The rest of the classrooms (4, 6, and 8) either did not provide any form of data or did not provide any form of explanation at all.

Reasoning. This component is critical since it is the one that legitimizes the claim by showing how the data support the claim, why it is relevant (Table 9). We focus on two aspects of reasoning: alignment between the data and the claims, and type of link (justification) provided. To judge the alignment we asked whether the evidence provided supported the claim stated. We considered three possibilities: (a) complete alignment – the evidence provided supported the claims stated, (b) partial alignment – the evidence provided only supported some of the claims stated, and (c) no alignment – the evidence provided did not support any of the claims stated or evidence was irrelevant.

As previously noted, few students provided explanations in their complete form (i.e., with the three components). Only these students were considered in judging alignment between claims and evidence. From these students, those from Classrooms 5 showed the highest percentage of complete alignment between the evidence provided and the claims

stated followed by partial alignment. In Classroom 7 it was just the opposite. The highest percentage was observed in partial assignment followed by complete alignment.

Table 9

Quality of Students' Reasoning by Characteristic and Classroom in Percentages

Item	Classrooms							
	1	2	3	4	5	6	7	8
Alignment of claim and evidence								
Completely	0.0	0.0	0.0	0.0	44.4	0.0	22.2	0.0
Partially	0.0	0.0	0.0	0.0	11.1	0.0	55.6	0.0
No	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hard to know[a]	11.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Not applicable[b]	66.7	100.0	100.0	77.8	11.1	33.3	22.2	88.9
No explanation	22.2	0.0	0.0	22.2	33.3	66.7	0.0	11.1
Type of link – connection of evidence to claim								
Elaborated connection	0.0	0.0	0.0	0.0	0.0	0.0	11.1	0.0
Simple connection	11.1	0.0	0.0	0.0	55.6	0.0	44.4	0.0
No connection	0.0	0.0	0.0	0.0	0.0	0.0	22.2	0.0
Not applicable[c]	66.7	100.0	100.0	77.8	11.1	33.3	22.2	88.9
No explanation	22.2	0.0	0.0	22.2	33.3	66.7	0.0	11.1

[a]Hard to know refers to cases in which additional information is required to evaluate the alignment of students' claim and the evidence provided, such as the context of the lesson or the terms used. [b]Claim or data only. [c]Claim and/or data only.

Basically, most of the students who provided a justification between their claims and their evidence used a simple connection with words such as "because" or "it is what we found in our data." Students did not elaborate the justification by describing or interpreting the data so that the link was more explicit. These results confirm Toulmin's (1958) contention that the weakest part of any explanation is the reasoning or what he called, warrant.

Extended Qualities of the Explanations

We focused on three aspects of the extended qualities, whether students considered: (a) issues related to the quality of the evidence or the reasoning (e.g., measurement error), (b) alternative explanations to respond the question at hand, and (c) implications of the findings (Table 10).

As expected, if explanations were provided, very few students addressed these three extended aspects in their explanations. Only in Classroom 7 did students focus on qualifying their explanations based on the quality of the evidence provided (66.7%). Most of these students mentioned human error as the main factor that could affect the quality of the evidence. None of the students in the sample considered alternative explanations.

Table 10

Extended Characteristic of Students' Explanations by Type and Classroom in Percentages

Item	Classrooms							
	1	2	3	4	5	6	7	8
Evaluation of quality of evidence or strength/ weaknesses of the link								
Discussion of strength of the reasoning (link)	0.0	77.8	77.8	0.0	66.7	0.0	22.2	0.0
Discussion of measurement error	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mention of strength of the reasoning (link)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mention measurement error	11.1	0.0	0.0	0.0	0.0	0.0	66.7	0.0
No evaluation	88.9	22.2	22.2	100.0	33.3	100.0	11.1	100.0
Implications of findings								
Connections beyond the curriculum	0.0	0.0	0.0	0.0	0.0	0.0	11.1	0.0
Connections within the curriculum	0.0	0.0	0.0	0.0	0.0	0.0	55.6	0.0
Application of the findings	0.0	0.0	0.0	0.0	0.0	0.0	11.1	0.0
Limitations of the findings	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
No implications	100.0	100.0	100.0	100.0	100.0	100.0	22.1	100.0

We considered four issues that students could include in their explanations when discussing the implications of their findings: (a) issues beyond the investigations (e.g., students discussed possible research questions or issues that can be studied based on the findings), (b) issues in other instigations within the unit (e.g., students discussed how Investigation 7 was connected to other investigations in FAST), (c) application of the findings (e.g., students discussed the real life application of the findings), and (d) limitations of the findings (e.g., students attended to the limitations or conditions when applying the

findings). Only in Classroom 7 did we observe students including implications of their explanations. The majority of the students (55.6%) made connections to other investigations within the unit.

Level of Students' Understanding

The score of the students' level of understanding was based on the trajectory previously explained (see Table 3). We created a 4-point scale score. When students noted that mass and volume were related (Level 4), the appropriate level at Investigation 7, we scored their performance as 2. If students' understanding was above Level 4 (i.e., Levels 5 or 6), we scored performance as 3. If students' understanding was at Level 3, we scored performance as 1. If students' understanding was at Levels 2 or 1, we scored performance as 0.

Table 11 provides information about the students' level of understanding observed in their explanations. Students in Classrooms 2, 5 and 7 showed the appropriate level of understanding in Investigation 7. Some students in Classroom 3 did not reach the appropriate level based on the analysis of their explanations. In Classroom 6, only a few students provided information (claims) on which to judge their level of understanding. The level of understanding reflected in their claims showed naïve or inappropriate conceptions.

Table 11

Descriptives of Students' Level of Understating by Classroom

			Classroom							
Item	Max	*n*[a]	1	2	3	4	5	6	7	8
Student understanding	3	9								
Mean			0.57	2.0	0.67	1.43	2.0	0.0	2.0	0.63
Standard deviation			0.79	0.0	1.15	0.98	0.0	0.0	0.0	0.52

[a]*n* per classroom.

Linking Quality of Students' Explanations to Their Performance

To explore the link we first provided information on explanation scores calculated for each explanation component as well as a composite score that involved the averaged sub-scores (Table 12). We then provided information on the students' performance across the four assessments used in the post-test. We also included a gain score ([post-test multiple-choice score] – [pre-test multiple-choice score]) and an achievement composite score in which the scores of all assessments are involved (Table 13). Finally, to link the quality of the

explanations with students' learning, we correlated their explanation scores with the different achievement scores.

Two classrooms showed consistently high mean scores across the different types of explanation scores, Classrooms 5 and 7, followed by Classrooms 2 and 3. The lowest scores are observed in Classroom 6. The mean scores of the other three classrooms can be located towards the lower end, rather than the higher end. It is important to note that the observed mean scores of the classrooms studied are not close to the maximum mean scores possible. Even in those classrooms, where the construction of explanations was practiced, the emphasis on transforming the data to evidence, and evidence to explanations was not a common practice in those classrooms. Despite the fact that the construction of explanations is a critical aspect of inquiry-based instruction, it is clear that the construction of explanation is not constant in every scientific-inquiry classroom. Furthermore, in those classrooms in which written explanations were found, the quality of students' explanations varied from classroom to classroom.

Table 12

Students' Explanation Scores

			Classrooms							
Type of score	Max.	n[a]	1	2	3	4	5	6	7	8
Type of explanation	3	9								
Mean			1.0	1.78	1.11	0.78	1.89	0.33	2.56	0.89
Standard deviation			0.87	0.44	0.33	0.44	1.45	0.50	0.88	0.33
Evidence	6	9								
Mean			0.67	0.78	3.33	0.0	3.22	0.0	3.44	0.0
Standard deviation			2.0	0.44	2.06	0.0	2.53	0.0	1.81	0.0
Focus	5	9								
Mean			0.57	3.78	1.67	2.86	4.86	0.0	4.38	0.88
Standard deviation			1.51	0.67	2.88	2.41	0.41	0.0	1.77	1.64
Alignment	3	9								
Mean			0.11	0.0	0.0	0.0	1.56	0.0	1.78	0.0
Standard deviation			0.33	0.0	0.0	0.0	1.50	0.0	1.09	0.0
Composite score	17	9								
Mean			2.22	6.33	5.00	3.00	8.89	0.33	10.56	1.67
Standard deviation			4.43	1.22	1.65	2.69	6.79	0.50	3.97	1.66

[a]*n* per classroom.

The issue then is to find out whether constructing explanations really matters in helping students move forward in their learning. Research in which diverse treatments have been implemented to help students to construct explanations has documented a positive impact on student learning (Bell & Linn, 2000; Boscolo & Mason, 2001; Keys, 2000). Is it reasonable to expect that this positive relation can also be observed in inquiry-based classrooms in which the practice of constructing explanations is naturally observed without any treatment implemented?

To address this question we carried out a series of simple correlations that could allow for determining the association between students' explanation scores and their performance on a series of end-of-unit assessments. First, we present the means and standard deviations on the assessments administered at the end of the 12 investigations (Table 13). Results indicate that students' mean performance by group varied across assessment. Although not always the same classrooms ranked the highest or lowest across assessments, overall they followed a similar pattern, which very much resembles the orders of students' explanations scores (see Table 12). In general, the highest mean scores across assessments were consistently observed for Classrooms 5 and 7, followed by Classroom 3. The lowest mean scores were observed, in general, for Classrooms 2 and 6.

Table 13

Descriptives of Students' Scores by Type of Assessment

Type of assessment	Max	n[a]	Classrooms							
			1	2	3	4	5	6	7	8
Multiple-choice post	43	9								
Mean			24.78	20.55	27.33	24.67	27.33	20.00	28.22	26.77
Standard deviation			8.55	7.07	7.65	6.26	8.39	9.06	8.59	7.84
Predict-observe-explain	7	9								
Mean			2.55	3.00	3.25	1.37	5.00	1.62	4.25	2.50
Standard deviation			2.30	2.07	2.65	1.19	3.16	1.30	2.29	1.69
Performance assessment	32	9								
Mean			18.12	11.75	16.66	17.50	18.50	12.57	20.78	16.33
Standard deviation			10.24	2.05	9.31	7.74	9.63	5.74	6.03	7.77
Short open-ended WTSF	6	9								
Mean			1.55	2.00	3.00	2.50	2.22	2.37	3.42	2.77
Standard deviation			1.58	1.65	0.50	1.69	1.39	1.19	0.53	1.20
Gain score (MC$_{post - pre}$)	43	9								
Mean			9.11	8.55	13.44	10.00	12.63	5.12	12.89	12.44
Standard deviation			6.41	7.66	5.17	5.24	7.22	5.64	7.44	7.68
Assessment composite	88	9								
Mean			48.62	37.50	49.87	46.00	65.60	38.42	61.71	49.62
Standard deviation			19.59	10.58	18.99	13.79	6.19	15.24	13.50	15.17

[a] n per classroom

The correlations observed are presented in Table 14. We focused on simple correlations that involved the eight groups since the pooled-within correlation may not provide an unbiased estimation due to the small sample size within each group and of classrooms. In the table, we also reported the means, standard deviations, and medians of simple correlation coefficients within classrooms, assuming that a strong relation can be observed within classrooms as well as when students were pulled together regardless of classrooms. This enables us to detect whether the reported overall correlations were artifacts of score differences between classrooms.

The pattern observed across the two sets of correlations indicates a significant positive relation between the quality of the students' explanations and students' performance at the

end of the investigations. However, it is important to notice that the magnitude of the correlations varied according to the type of assessment at hand.

Table 14

Correlations between the Explanation Composite Score and the Diverse Assessments

		Explanation composite score[a]			
		Overall simple correlation	Descriptives of simple correlations within classrooms		
Assessment	*n*		Mean	*SD*	Median
Multiple-choice post	65	0.255*	0.207	0.304	0.281
Predict-observe-explain	56	0.349**	0.170	0.250	0.194
Performance assessment[b]	59	0.338**	0.386	0.252	0.393
Short open-ended (WTSF)[b]	61	0.277*	0.295	0.353	0.333
Gain score of multiple-choice	63	0.256*	0.188	0.443	0.241
Assessment composite score[b]	53	0.472**	0.407	0.339	0.520

[a]Seven students were identified as outliers and were not considered in the analyses. Simple correlations when all students were included were all positive and lower in magnitude, but only two of them were still significant (POE and composite score). Those students were excluded from further analysis in Table 15. [b]Assessments yielded strong correlations with students' explanations composite scores.
*Correlation significant at 0.05. **Correlation significant at 0.01.

The magnitudes of standard deviations of the correlation coefficients (ranging from .250 to .443) suggest a strong effect of between-classroom variation. Most likely, depending on the scaffolding or constraints provided by teachers for students to construct their explanations, students may or may not be supported to demonstrate the same level of understanding showed in the set of assessments administered. These assessments were standardized testing situations with the same supports in the assessment tasks using carefully designed items to sufficiently differentiate students' understanding. In contrast, notebooks were used by teachers with varying types of prompts and scaffolding, which might lead students to reflect different aspect and level of understanding.

The highest correlation was observed between the explanation score and the composite score among all the assessments since it represented the summarized information across the four assessments. The results for overall sample and within classrooms show that performance assessments and short open-ended item (WTSF) were highly correlated with explanation scores compared to the multiple-choice post-test scores and gain scores. These results seem to suggest that the nature of the assessment may be related to the magnitude of this relationship pattern. For example, in multiple-choice tests students are required to select

a response, whereas in the performance assessment, students are required to explain. Therefore, the nature of this assessment seems to be more aligned to what students need to do in constructing explanations. We interpreted these results as an indication that engaging students in the construction of explanations is likely to have a positive impact in students' learning and achievement of the content.

Based on these results we asked whether there was an explanation component (i.e., claim or evidence) that could be considered critical in helping students to improve their learning. To answer this question we conducted another series of correlational analyses with the explanation sub-scores (Table 15). Overall, students' explanations on the two sub-aspects were positively correlated with the six types of post-test scores, mostly with a small or medium effect size. Scatter plot analysis confirmed that several cases of weak correlation within classrooms, indicated by the descriptive statistics of correlation coefficients, were solely due to lacking score variation in outlier classrooms. In general, we think that it is possible to claim that focusing on at least one of the explanation components, claim or evidence, seems to have a positive impact on students' performance. This probably explains why students' performance in Classroom 3 was higher than in other classrooms. In this group, students were encouraged to focus on patterns of data. This information is important since explanations are, in the end, an account for patterns of data (Sandoval, 2001).

Table 15

Correlations between the Types of Explanation Scores and the Diverse Assessments

Type of assessment	Claim focus score					Quality of evidence score				
		Overall simple correlation	Descriptives of simple correlations within classrooms				Overall simple correlation	Descriptives of simple Correlations within classrooms		
	n		Mean	*SD*	Median	*n*		Mean	*SD*	Median
Multiple-choice post	46	0.242	0.031	0.545	0.002	65	0.226	0.052	0.269	-0.046
Predict-observe-Explain	41	0.339*	-0.143	0.379	-0.106	56	0.397**	0.223	0.092	0.196
Performance assessment	44	0.362*	0.143	0.474	0.020	59	0.221	0.157	0.332	0.225
Short open-ended (WTSF)	43	0.198	0.214	0.282	0.227	61	0.227	0.384	0.334	0.333
Gain score	45	0.252	0.179	0.539	0.200	63	0.282*	0.160	0.226	0.166
Assessment composite score	41	0.403**	-0.038	0.497	-0.157	53	0.361**	0.249	0.374	0.467

*Correlation significant at 0.05. **Correlation significant at 0.01.

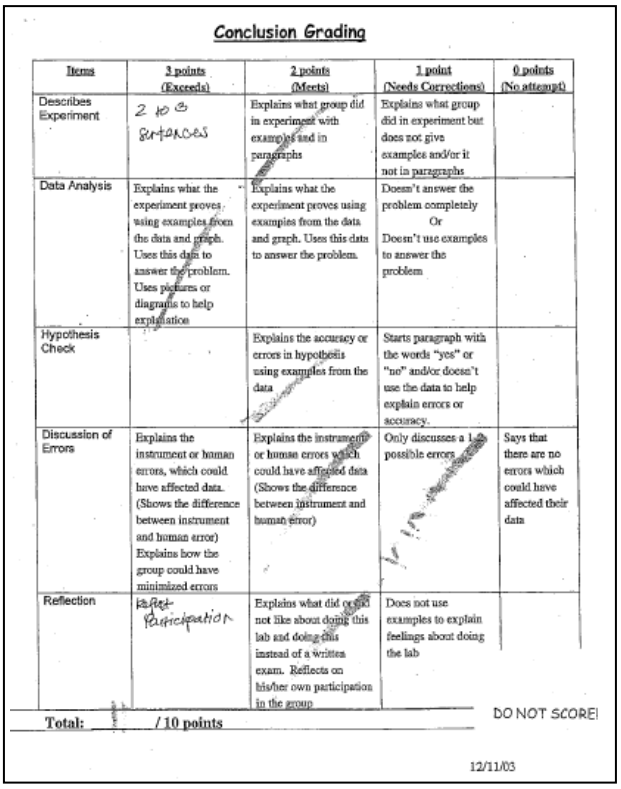
Finally, we attended to the relation between students' level of understanding about sinking and floating reflected in their explanations and the end-of-unit assessments. Because of the lack of variation in the students' level of understanding score, it was not a surprising finding that the magnitude of the correlations between the level of understanding score and the students' performance score with the diverse assessments were all positive, but only one was significant: the correlation with the POE assessment (r = .38; p = 0.01). The positive correlation indicates that it is possible to obtain information about students' understanding from what they write in their notebooks. This finding confirms previous results on the use of notebooks as assessment tools (Ruiz-Primo & Li, 2004; Ruiz-Primo, Li, Ayala, & Shavelson, 2004).

A Quick Look at the Characteristics of the Instructional Prompts

As mentioned previously, different prompts have been studied to support the construction of explanations, ranging from generic skeletons to investigation-based questions. We captured the general characteristics of the prompts found in the sample of classrooms in trying to find some general characteristics that may seem to help students in constructing explanations. In this section we present information on the characteristics of the instructional prompts. We created nine categories of prompts (e.g., teacher questions, conclusion section, and skeleton with specific questions).

Across the entire set of classrooms, the type of prompts with the highest percentage was a format with teacher questions (20.8%), followed by the FAST Summary questions (18.1%) and a conclusion section (13.9%).

We focused on the characteristics of the prompts in the classrooms in which the highest scores in the explanation were observed (Classroom 5 and 7). In Classroom 5, the teacher provided students with a skeleton of a report format. The format already included printed information about the purpose, background, and procedure. The teacher provided students with investigation-based questions that guide them to interpret the data step by step (e.g., *If the mass of a floating object is known, can the volume of displaced water be predicted? Be sure to cite your data that supports your answer.*). In Classroom 7 the explanations were found in a conclusion section. The teacher provided students with a piece of paper with scoring criteria of the information to be included in the conclusion section (see Figure 4). Some students had this piece of paper taped to their notebooks. It is important to notice the different aspects considered by the teacher and the criteria used for scoring (e.g., discussion of errors).



Conclusion Grading

Items	3 points (Exceeds)	2 points (Meets)	1 point (Needs Corrections)	0 points (No attempt)
Describes Experiment	2 to 3 sentences	Explains what group did in experiment with examples and in paragraphs	Explains what group did in experiment but does not give examples and/or it not in paragraphs	
Data Analysis	Explains what the experiment proves using examples from the data and graph. Uses this data to answer the problem. Uses pictures or diagrams to help explanation	Explains what the experiment proves using examples from the data and graph. Uses this data to answer the problem.	Doesn't answer the problem completely Or Doesn't use examples to answer the problem	
Hypothesis Check		Explains the accuracy or errors in hypothesis using examples from the data	Starts paragraph with the words "yes" or "no" and/or doesn't use the data to help explain errors or accuracy.	
Discussion of Errors	Explains the instrument or human errors, which could have affected data. (Shows the difference between instrument and human error) Explains how the group could have minimized errors	Explains the instrument or human errors which could have affected data (Shows the difference between instrument and human error)	Only discusses a 1-2 possible errors	Says that there are no errors which could have affected their data
Reflection	Self Participation	Explains what did or did not like about doing this lab and doing this instead of a written exam. Reflects on his/her own participation in the group	Does not use examples to explain feelings about doing the lab	

Total: ____ / 10 points DO NOT SCORE!

12/11/03

Figure 4. Scoring sheet provided in Classroom 7 to guide students in the conclusion writing.

The prompts had a common characteristic with those used in some studies previously cited, a *guided* support for students to construct explanations. The one from Classroom 5 is more specific than the one from Classroom 7. The main difference between the two is that the former is content-based and the later generic, another characteristic observed in previous studies.

It is important to mention that the curriculum summary questions used by three of the teachers in the sample (Classrooms 4, 6, and 8) focused on the claims only (i.e., What relationship, if any, are there between the mass of an object and the volume of water displaced?). This might be the main reason why complete explanations were absent in these classrooms. The difference between Classrooms 4 and 8, and Classroom 6 is that in the first two, the curriculum summary questions were supplemented with additional questions, but not focused on the explanation issues.

We concluded that an appropriate practice for supporting students in the construction of explanations is to provide them with specific prompts with moderate guidance, rather than general and minimum guidance (Aschbacher & Alonzo, 2006). Too much or too little scaffolding from the teachers leads to a lack of variation in students' explanations and

therefore, to an undifferentiated level of understanding (e.g., all students provided two examples as evidence because of the explicit requirement from the teacher). Needless to say, the role of the teacher in supporting students' construction of explanations should be considered critical too. Further research needs to be done around types and extent of teacher supports which would lead to desirable learning outcomes and assessment information.

Conclusions

In this study, we collected the students' science notebooks in eight, middle school, inquiry-based science classrooms. We analyzed them to find out how common it was for middle school students to write explanations using the data they collected during a physical science investigation.

When explanations were found, we analyzed their quality and we linked it with students' performance in diverse assessments focusing on the content studied. We also analyzed the characteristics of the instructional prompts used to engage students in constructing explanations to determine what characteristics seem to be linked to explanations of high quality.

We found that only 18% of the notebooks analyzed had explanations with the three expected components, a claim, evidence to support it, and a reasoning that justified the link between the claim and evidence. The majority (40%) of the "explanations" found were presented as claims only, without any supporting data from the investigation that students carried out. We also observed that some teachers' students gave priority to evidence by describing data patterns in the conclusion section (Classroom 3). However, the data were not utilized as evidence to construct or support claims, at least not in a written form.

We believe that scientific inquiry is fundamentally about collecting data, transforming that data into evidence, and that evidence into explanations (Duschl, 2003; Sandoval & Reiser, 2004). Furthermore, if students do not assimilate evidence in ways that question their current understanding, it is very unlikely that conceptual change can take place (Sandoval, 2001). However, the findings suggest that the instructional practice of constructing explanations has not been widely implemented despite its significance in the context of inquiry-based instruction. Moreover, results indicate a great variation in teachers' implementation across classrooms. These results raised essential issues around scientific inquiry-based instruction: How much do we know about whether or not a fundamental premise of scientific inquiry instruction is being met in everyday science classrooms? What is the impact on students' learning and their understanding of the nature of science if this premise is not met?

Why do students omit constructing explanations based on the data collected? Why do students think that the data collected constitute answers in themselves? Our results lead to question whether it is appropriate to assume that students are the ones who ignore seeking patterns or contradictory data (Sandoval, 2001). It seems to be possible to hypothesize that what students are missing and lacking is experience and guidance in the fundamental activities of constructing explanations. Without such experience and guidance, it is not surprising to find that constructing explanations is challenging for students (Kuhn & Reiser, 2004).

Furthermore, flaws in students' explanations, such as citing irrelevant or insufficient data to support a claim, may not result from an unscientific way of thinking. This may be related to the lack of opportunities to be engaged in scientific practices where they develop, argue, and evaluate explanations through their own investigations. Constructing explanations should be seen as the means to understand a phenomenon and as a means to get engaged in the inquiry process (Kuhn & Reiser, 2004).

Another important finding in this study is the link between the quality of the students' explanations and their performance in diverse assessments administered at the end of the 12 FAST investigations. The magnitudes of the correlations indicate that engaging students in the construction of high quality explanations might be related to higher levels of student performance. We interpreted this finding as evidence that engaging students in the construction of explanations can lead to expected positive impact on their understanding. The opportunities to construct these explanations, however, seem to be limited.

Finally, the study also provides evidence on the technical qualities of students' science notebooks as assessment tools. We provided evidence about the consistency among raters to code and score explanations, and students' level of understanding using the approach proposed. However, we could not provide conclusive evidence of its validity, since notebook explanation scores and sub-scores were only positively correlated with some of the post-test assessments implemented. We believe that in order to ensure the technical quality of notebook scores, notebooks need to be implemented in certain ways. These ways should allow students to reflect more accurately their level of understanding, and teachers to reflect reform-oriented practices that are more consistent from lesson to lesson (e.g., Aschbacher & Alonzo, 2006) and from direct observations to notebooks (see Borko, Stecher, Alonzo, Moncure, & McClam, 2005).

We are convinced that the use of prompts with adequate level of guidance is what is needed to collect information that can be used for assessment purposes, both formative and

summative. All the teachers scored in this study had notebook prompts with diverse degrees of guidance. The ones that were the best for instructional and assessment purposes seem to be those that scaffold students to provide pieces of information relevant to the explanations (e.g., evidence, provide the relationship between variables), but that also allowed students to do their own thinking. Those that were of low guidance (like the one used in Classroom 1) provided structure (a skeleton with subtitles such as conclusion), but with an insufficient focus. The prompt with high level of guidance (like the one in Classroom 2) promoted copying from the board and did not allow student to incorporate their own thinking. In sum, all teachers used prompts or templates of different quality that allowed or hindered students to make explicit their level of understanding. What we think is missing is aid to teachers in understanding the purposes of using notebooks and prompts that are of the "just right" guidance (Aschbacher & Alonzo, 2006) and that focus on critical aspects of learning. For example, a prompt just asking students for a claim and evidence with a low level of guidance (i.e., My Claim: _____, and My Evidence: _____) elicits vague student responses that are not conceptually focused. Just right guidance seems to involve more questions that can guide students' own thinking (e.g., explanations are constructed around a set of coherent questions that students need to respond in relation to the components of explanations; Sandoval, 2004).

Finally, we need to learn more about the nature of the challenges that students face when constructing and communicating explanations, assuming that appropriate opportunities are provided for doing so. Designed supports are available, but how can they be scaled up? How can curriculum materials support this process in a better form? We need to find ways to improve teachers' practices in implementing scientific inquiry-based curricula.

References

Alonzo, A. C. (2001). *Using student notebooks to assess the quality of inquiry science instruction*. Paper presented at the American Educational Research Association annual meeting. Seattle, WA.

Applebee, A. N. (1984). Writing and reasoning. *Review of Educational Research*, *54*, 577-596.

Applebee, A. N. (1997). Analysis and description of students' learning during science classes using constructivist-based model. *Journal of Research in Science Teaching*, 34, 303-318.

Aschbacher, P. & Alonzo, A. (2006). Examining the utility of elementary science notebooks for formative assessment purposes. *Educational Assessment*, *11*(3-4), 179-203.

Bass, K. M., Baxter, G. P., & Glaser, R. (2001). *Using reflective writing exercises to promote writing-to-learn in science*. Paper presented at the American Educational Research Association annual meeting, Seattle, WA.

Baxter, G. P., Bass, K. M., & Glaser, R. (2000). *An analysis of notebook writing in elementary science classrooms*. CSE Technical Report 533. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing/Graduate School of Education & Information Studies. University of California, Los Angeles.

Bell, P. & Linn, M. C., (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education*, *22*, 797-817.

Bereiter, C. & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Borko, H., Stecher, B. M., Alonzo, A., Moncure, S., & McClam, S. (2005). Artifact packages for charactering classroom practice: A pilot study. *Educational Assessment*, *10*(2), 73-104.

Boscolo, P. & Mason, L. (2001). Writing to learn, writing to transfer. In P. Tynjälä, L. Mason, & K. Lonka (Eds.), *Writing as a learning tool. Integrating theory and practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Brown, A. L. & Campione, J. C. (1990). Communities of learning and thinking, or a context by any other name. *Human Development, 21*, 108-125.

Chi, M. (2000). Self-explaining expository texts. The dual process of generating inferences and repairing mental models. In R. Glaser (Ed.). *Advances in instructional psychology* (Vol. 5, pp. 165-238). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Duschl, R. A. (2003). Assessment of inquiry. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41–59). Arlington, VA: National Science Teachers Association Press.

Halliday, M. A. K. & Martin, J. R. (Eds.). (1993). *Writing science. Literacy and discursive power*. Pittsburgh, PA: University of Pittsburgh Press.

Hand, B. & Prain, V. (2006). Moving from border crossing to convergence of perspectives in language and science literacy research and practice. *International Journal of Science Education*, *28*(2-3), 101-107.

Kelly, G. J., Drucker, S., & Chen, K. (1998). Students' reasoning about electricity: Combining performance assessments with argumentation analysis. *International Journal of Science Education*, *20*(7), 849-871.

Kenyon, L. & Reiser, B. J. (2006). *A functional approach to nature of science: Using epistemological understandings to construct and evaluate explanations*. Paper presented at American Educational Research Association annual meeting. San Francisco, CA.

Keys, C. W. (2000). Investigating the thinking processes of eight grade writers during the composition of a scientific laboratory report. *Journal of Research in Science Teaching*, *37*(7), 676-690.

Keys, C. W., Hand, B., Prain, V., & Collins, S. (1999). Using the science writing heuristic as a tool for learning from laboratory investigations in secondary science. *Journal of Research in Science Teaching*, *36*(10), 1065-1084.

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review*, *11*(3), 203-270.

Klein, P. D. (2004). Constructing scientific explanation through writing. *Instructional Science*, *32*, 191-231.

Klein, P. D. (2006). The challenges of scientific literacy: From the viewpoint of second-generation cognitive science. *International Journal of Science Education*, *28*(2-3). 143-178.

Kuhn, L. & Reiser, B. J. (2004). *Students constructing and defending evidence-based scientific explanations*. Paper presented at the NARST Annual Meeting. Dallas, TX.

Kuhn, L. & Reiser, B. J. (2006). *Structuring activities to foster argumentative discourse*. Paper presented at American Educational Research Association annual meeting. San Francisco, CA.

Martin, J. R. (1993a). Literacy in science: Learning to handle text as technology. In M. A. K. Halliday & J. R. Martin (Eds.). *Writing science. Literacy and discursive power* (pp.166-202). Pittsburgh, PA: University of Pittsburgh Press.

Martin, J. R. (1993b). Life as a noun: Arresting the universe in science and humanities. In M. A. K. Halliday & J. R. Martin (Eds.). *Writing science. Literacy and discursive power* (pp.221-267). Pittsburgh, PA: University of Pittsburgh Press.

McNeill, K. & Krajcik, J. (2006). *Supporting students' constructions of scientific explanation through generic versus context-specific written scaffolds*. Paper presented at the American Educational Research Association annual meeting, San Francisco, CA.

National Research Council. (1996). *The National Science Education Standards.* Washington, DC: National Academy Press.

National Research Council (2000). *Inquiry and the National Science Education Standards*. Washington, DC: National Academy Press.

Pottenger, F. & Young, D. (1992). *The Local Environment: FAST 1 Foundational Approaches in Science Teaching*. University of Hawaii at Manoa: Curriculum Research and Development Group.

Prain, V. (2006). Learning from writing in secondary science: Some theoretical and practical implications. *International Journal of Science Education*, *28*(2-3), 179-201.

Rivard, L. P. & Straw, S. B. (2000). The effect of talk and writing on learning science: An exploratory study. *Science Education*, *84*, 566-593.

Ruiz-Primo, M. A. & Li, M. (2004). On the use of students' science notebooks as an assessment tool. *Studies in Educational Evaluation*, *30*, 61-85.

Ruiz-Primo, M. A., Li, M., Ayala, C. C., & Shavelson, R. J. (1999, March). *Student science journals and the evidence they provide: Classroom learning and opportunity to learn.* Paper presented at the meeting of the National Association of Research in Science Teaching, Boston, MA.

Ruiz-Primo, M. A., Li, M., Ayala, C. C., & Shavelson, R. J. (2004). Evaluating students' science notebooks as an assessment tool. *International Journal of Science Education*, *26*(12), 1477-1506.

Ruiz-Primo, M. A., Li, M., & Shavelson, R. J. (2001). *Looking into students' science notebooks: What teachers do with them?* CSE: Technical Report 562. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing/ University of California, Los Angeles.

Sandoval, W. (2001). *Students' uses of data as evidence in scientific explanations*. Paper presented at the American Educational Research Association annual meeting. Seattle, WA.

Sandoval, W. (2003). Conceptual and epistemic aspects of students' scientific explanations. *The Journal of the Learning Sciences*, *12*(1), 5-51.

Sandoval, J. H. (2004). Evaluation issues and strategies in consultee-centered consultation. In N. M. Lambert, I. Hylander & J. H. Sandoval (Eds.), *Consultee-centered consultation: Improving the quality of professional services in schools and community organizations* (pp. 391-400). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Sandoval, W. (2005). *Ability, ontology, and method: A commentary on Hammer, Russ, Mikeska, and Scherr*. Inquiry Conference on Developing a Consensus Research Agenda. Piscataway, NJ.

Sandoval, W. & Reiser, B. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*, 345-372.

Shavelson, R. J. & Young, D. (2000). *Embedding assessments in the FAST curriculum: On the beginning the romance among curriculum, teaching and assessment*. Proposal submitted at the Elementary, Secondary and Informal Education Division at the National Science Foundation.

Shepardson, D. P. & Britsch, S. J. (1997). Children's science journals: Tools for teaching, learning, and assessment. *Science and Children*, *34*(5), 13-17, 46-47.

Simon, S., Erduran, S., & Orborne, J. (2006). Learning to teach argumentation: Research and development in the science classroom. *International Journal of Science Education*, *28* (2-3), 235-260.

Stanford Education Assessment Laboratory (2003). *FAST Teacher's Guide to the Reflective Lessons*. Stanford, CA: Stanford University.

Tishman, S. & Perkins, D. (1997). The language of thinking. *Phi Delta Kappan*, 78, 368-374.

Toulmin, S. (1958). *The uses of arguments*. Cambridge, England: Cambridge University Press.

Tzou, C. (2006). *Characterizing teachers' support of constructing scientific explanations from a discourse perspective*. Paper presented at the American Educational Research Association annual meeting, San Francisco, CA.

Warren-Little, J., Gearhart, M., Curry, M., & Kafka, J. (2003). Looking at student work for teacher learning, teacher community, and school reform. *Phi Delta Kappan*, *85*(3), 185-192.

Yin, Y. (2005). *The Influence of Formative Assessments on Student Motivation, Achievement, and Conceptual Change.* Stanford University, Stanford, CA.