How Good Is This Evidence?
Students' Epistemic Competence in Evidence Evaluation

Clark A. Chinn, Leah C.-C. Hung, Randi M. Zimmerman, Ravit Golan Duncan
Graduate School of Education, Rutgers University, 10 Seminary Place, New Brunswick, NJ 08901
clark.chinn@gse.rutgers.edu, randi.zimmerman@gse.rutgers.edu, leahcchung@gmail.com,
ravit.duncan@gse.rutgers.edu

Abstract: Inquiry environments in science classes have increasingly incorporated more features of authentic scientific practice. However, relatively few environments have incorporated a critical feature of scientific practice: evaluation of evidence quality. This paper reports results from two studies in middle-school life science classes that investigate seventh graders' competence in evaluating evidence. Overall, we found evidence that students have strong evaluative capabilities that can be built upon in instruction.

Introduction: Integrating Evidence Evaluation into Inquiry Environments

Chinn and Malhotra (2002) argued that, for inquiry-oriented instruction to be effective in improving students' reasoning, it is important that these environments be as authentic as possible. Oversimplified inquiry environments can promote oversimplified student epistemologies that are counterproductive in the complexity of real-life settings. For example, when students learn to control variables in oversimplified laboratory tasks, they may come to think of scientific inquiry as simple and algorithmic, just a matter of checking off a set of provided variables, failing to appreciate the difficulties of deciding what variables to control in real-life inquiry, devising how to control them, and so on. Chinn and Malhotra (2002) touted more complex inquiry environments such as the BGuILE environment (Reiser et al., 2001) that afforded students the opportunities to reason in more realistic ways about more complex problems of the sort that they will encounter out of school.

However, there is one important aspect of authentic scientific inquiry that remains relatively infrequent in inquiry environment designs, especially those developed for younger students such as middle-school students. Specifically, in authentic science, scientists not only use evidence to evaluate theories; they also evaluate the evidence itself. They consider how strongly to weight the evidence based on their evaluations of evidence quality (e.g., whether a study has an adequate sample size, properly controlled variables, valid and reliable measures, and so on) (Staley, 2004). They appropriately discount low-quality evidence. But most inquiry environments have not featured low-quality evidence with the intent of helping students learn how to reason about such evidence. There are plausible arguments for holding off on introducing extensive discussions of methodological features of studies. Students may have difficulty learning to identify methodological flaws; if this is so, then there is a danger that students will draw incorrect conclusions from evidence. For example, a student presented with methodologically flawed evidence against global warming might fail to notice its flaws and use the evidence as a centerpiece of an argument that global climate is not warming. Given worries such as these, there is a need for research that investigates students' abilities to evaluate evidence quality, as a first step in understanding how methodological diversity in evidence can be introduced to inquiry environments.

There is relatively limited research on how students evaluate the methodological processes used by scientists (Allchin, 2011). Some research has examined students' understanding of error in simple scientific experiments (Masnick & Klahr, 2003), but there is much less work examining evaluation of more complex, realistic studies. Those educational studies that provide students with details of research reports often simply code whether participants refer to "methods," and do not provide more detailed analyses of what exactly participants have to say about such methods. Our research attempts to help us begin to build an understanding of the strengths and weaknesses of students' reasoning about evidence quality.

In this paper, we report the results of two exploratory investigations of students' skill in evaluating evidence quality in science inquiry environments. This is part of a larger project investigating students' ability to learn to coordinate theories and evidence while simultaneously evaluating evidence quality. In the spirit of the "Reports & Reflection" category of ICLS papers, we report some critical initial findings that can provide grounding for future design work. We view this work as an essential first step in developing inquiry environments that are effective at promoting the critical epistemic practice of evaluating evidence quality.

Context of Research: PRACCIS

The context of our research is the PRACCIS (Promoting Reasoning and Conceptual Change in Science) project (Chinn, Duncan, Dianovsky, & Rinehart, 2013). In this model-based inquiry project, students use evidence to develop and revise models, as well as to decide between two or more models. We use a variety of scaffolds to promote thinking about theories and evidence, including the MEL (Model-Evidence Link) diagram (Rinehart, Duncan, & Chinn, in press), in which students use arrows of different types to indicate whether evidence supports, strongly supports, contradicts, strongly contradicts, or is irrelevant to one or more models. Like most other inquiry environments, our early implementations of PRACCIS used only “good” evidence that students were expected to take seriously as they developed models. Previous to the studies reported in this paper, we made no attempt to design “bad” or low-quality evidence into the curriculum so that students could learn to reason about such evidence. The present paper outlines our exploratory to work to authentically incorporate “bad” evidence into our curricula so that students have an opportunity to learn to evaluate evidence.

Study 1: Initial Explorations Using an Inquiry Unit with Low-Quality Evidence

Our initial exploration developing inquiry units with low-quality evidence occurred late in a school year after students had engaged in model-based inquiry for over 7 months. We developed a food web unit in which seventh graders used evidence to choose between two alternative models of the food web of an arctic habitat involving seals, foxes, polar bears, salmon, and plankton. We employed evidence of different quality, ranging from low-quality to high-quality evidence, to encourage student reflection and discussion about evidence quality. Two teachers implemented this unit in their classes.

In one low-quality piece of evidence, a hotel owner interviewed on a simulated Alaskan radio show reported that 350 guests at her hotel reported seeing 11 arctic foxes this year and no salmon in the ocean, but the guests a year ago reported seeing 16 arctic foxes and two salmon. This evidence involved haphazard, unrepresentative samples both of guests and observations. A second piece of evidence showed four photos snapped by a fisherman, each showing one or two polar bears eating a seal. Other evidence involved more systematic observations and population counts of different species.

We wondered whether students would detect the poor quality of some of the evidence without extensive coaching from the teacher, or indeed without being told in any way that any evidence was problematic. The following transcript excerpts present one class’s first discussion of these issues:

Teacher: So what does this evidence tell us about how many foxes and salmons were in the area this year and last year? Blair?
Blair: Uh, that there are more arctic foxes living uh in the area near the hotel but there was more the year before because it uh went down the next year. And the salmon, there aren't a lot in the ocean or there aren't a lot around the ocean where people are looking.
Teacher: [to the class] You need to listen to him, and you need to respond to him. I'm not saying anything.
Regina: Um, well, um, what I noticed in this evidence is where there are less arctic foxes, um, there weren't any, um, no one saw any salmon.
Teacher: No, no, just look at this information right here and say, what does the evidence show.
Regina: Um well, the less arctic foxes there were, there weren't any salmon and the less arctic foxes there were a few more. And um, I conclude from this evidence that um, if there is less arctic foxes there is, um, more ring seals. And where are there more ring seals, …..

To this point, there was no hint that any students recognized that the evidence might be of such low quality that it should be ignored or severely discounted. Instead, these two students (Blair and Regina) seem to have assumed that the evidence was good evidence, and that they needed to account for this evidence when choosing which food web was better. However, as the discussion continued (below), two students--without any hints or coaching by the teacher--commented on problems with the evidence.

Teacher: Okay. Anybody want to say anything to her? Yes, go on.
Andy: Well it's not really easy to see salmon in the ocean either, because they're underwater and stuff. So, you'd probably be pretty lucky to see any so just ’cause somebody saw two last year and there are more arctic fox don't mean that because there are more fox. There would be more salmon. ’Cause doesn't arctic fox eat salmon anyway?
Teacher: Did anyone have some thoughts about what she was saying?
Candace: Um, in the ocean, and they say that they saw two fish or whatever number they saw. If there's two, two fish in the water, there are more fish in the water because like how is there just going to be two fish? Like, that just makes no sense.
Teacher: Okay so what she was saying. So they start to give you up and down. Right? What was going up and what was going down. But does this tell us what was really happening with all the salmon?
Candace: No.
Teacher They're reading beyond that. They're now, Andy and Candace are making, critiquing; they are critiquing evidence rather than trying to draw a conclusion. Yes? And what else?

In this excerpt, Andy spontaneously questioned whether hotel visitors can reliably spot salmon in the ocean (thus noticing the possibility of error in observations). Candace added that the number of salmon that can be seen tells you little about the salmon that cannot be seen; in effect, Candace has detected a problem with inferring population sizes from a limited sample of observations. Only at this point did the teacher explicitly articulate that Andy and Candace are critiquing the evidence rather than trying to draw conclusions from it.

A few turns later, James argued that the "fox population is going down," and he offered this reason:

James: Because uh, last, because last March, there were only 11 foxes sighted. But the, but the year before there were a total of 16 sighted.

But other students responded to James by continuing their critique of the evidence:

Teacher: Sighted, okay. Yes, um, Erica?
Erica: That really doesn't tell you anything though, because it's just guesses. Like walking around, there, there's not like actual people going out trying to find these things. And it might not even be an arctic fox. They might just think it's an arctic fox.
Brianna: Also maybe like the arctic fox is kind of like, um, hibernated to another part of the forest, and they're not near the hotel anymore.
Teacher: You had your hand up?
Natalie: I was going to say that also, yeah, people could like not be looking for them or mistake them for an arctic fox which cannot be an arctic fox because they may not really know.
Regina: Um, well I don't think arctic foxes um will um intentionally um show themselves to people. So if people only saw 11, maybe there are more. They just saw it, there doesn't have to be 11. It could be um anywhere else. And….
Teacher: Keep going.
Regina: So um, if you just saw it, that doesn't mean there are only 11 there. Because I don't think arctic foxes will just show themselves.

Thus, students argued strongly against James's argument that the number of foxes sighted provides information about the actual fox population in this habitat. In addition to Andy's and Candace's earlier reasons, the new reasons given by the students include: (1) walking around without the intent to observe carefully is unreliable (Erica and Natalie); (2) observers might be mistaken in their observations due to poor ability to classify species seen (Erica and Natalie); (3) hotel guests might look in the wrong places, with foxes being elsewhere (Brianna); and (4) human observers underestimate the number of foxes because foxes hide from people (Regina). (Later in the discussion, a comment by James indicated that he accepted these criticisms as valid.)

Collectively, the students accurately (and in our estimation, impressively) identified a broad range of flaws with this evidence. The teacher did little to guide this discussion. Later, as the discussion continued, and students continued to give reasons to doubt the evidence, the teacher asked if the hotel guests' observations gave an accurate population count, and many students answered in chorus, "No." Thus, through the discussion, many students appear to have concluded that this evidence was poor quality and should be discounted.

This discussion and similar observed discussions from Study 1 provided encouraging preliminary evidence that seventh graders can reason effectively about authentic evidence sets that include low-quality as well as high-quality evidence. We explored students' capabilities in more detail in a second study.

Study 2: Seventh Graders' Evaluations of Evidence Quality

Encouraged by our experience with the food web unit, we developed a set of model-based inquiry units that systematically encouraged students to reflect on the quality of evidence. As students developed, revised, and evaluated models based on evidence, they also evaluated the evidence itself. In the analyses reported here, we focus on students' evidence evaluations in a six-week unit on organelles.

Method

We carried out a yearlong study involving over 500 seventh grades in 20 New Jersey science classes taught by 5 teachers implementing a life-science model-based inquiry curriculum that we designed. Two classes taught by each teacher were video recorded throughout this time. Our analyses draw on written data provided by 38 students in two of these teachers' video-recorded classes.

The students engaged in a variety of inquiry activities designed by the research team in extensive, iterative collaboration with participating teachers. The units engaged students in reasoning about evidence of widely varying quality. The data we discuss in this paper are all from an early unit in the curriculum, the organelles unit, in which students used evidence to develop a model of what chloroplasts do, used evidence to determine which of two models of mitochondria function is better (mitochondria produce "energy" versus mitochondria produce "movement") and used evidence to decide which of two models of nuclear function is better. In each unit, students were encouraged to reflect on and evaluate evidence quality. For example:

Chloroplast observations. After observing chloroplasts through a microscope, students considered how well their observational evidence could be taken to support any conclusions that they wanted to draw.

Hamster evidence. In the mitochondria lesson, students read a multimedia blog by a hamster owner (a college student) in which he performed microscopic observations of his dead hamster to determine whether the hamster's chronic low energy was caused by having too few mitochondria. The blog related several methodological problems experienced by the hamster owner as he carried out his investigation in a university lab.

Flagella evidence. In the mitochondria lesson, students learned about another study on another student's blog about observations of single-cell organisms with flagella. Several daily blog entries by the student provided a lively description of the procedures and findings of the study (more mitochondria were clustered near the flagellum than in other parts of the cell), but the results from only two cells were shown.

Cat evidence. The nucleus unit included a computer-based animation showing a study in which nuclei from glowing jelly fish were transplanted into cats, which produced the same protein that made the jellyfish glow, and yielded glowing cats. This was a well conducted study, albeit with a small sample size.

Students used these and other evidence to make their model-based judgments in each lesson. At several points, they also answered explicit questions about the quality of evidence. For example, students responded to this prompt about the cat evidence: "How good or bad is Evidence #2 [the cat evidence]? Write your reasons for your answer. Write to someone who might disagree with you." Our analyses focused on the students' responses to questions of this sort about the hamster evidence, the flagella evidence, and the cat evidence.

Research Questions

To support improved design of inquiry environments, we sought to understand the strengths and weaknesses of students' evaluation of evidence. The main questions were: (1) What is the range of students' responses when they are asked to evaluate evidence quality? (2) More specifically, how sensitive are students to methodological strengths and weaknesses in evidence? Given that scientists' evaluations of evidence quality typically focus strongly on how reliable the methodological processes were in the evidence, we wondered particularly whether students would pick up on reliable and unreliable methodological processes described in the evidence.

Coding

Two coders coded all of the data along two primary dimensions, discussing to reach agreement on all items. The first dimension captured the different categories of response to the data. The resulting categories from this dimension are presented in Table 1. The second dimension addressed how elaborated the responses were, on a 1 to 3 scale. Responses at Level 1 briefly mentioned features of the study but did not explain why those features made the evidence good or bad. Level 2 responses provided somewhat more detail than Level 1 responses. Level 3 responses pointed out specific features of the evidence and explained how these features supported an evaluation of the evidence as good or bad.

Table 1: Students' responses to prompts to evaluate how good the study is

Category	Example	% of students who gave response		
		Hamster evidence	Flagella evidence	Cats evidence
Student comments on study *results*.	It is OK evidence because it does sound like a low number of mitochondria.	14%	14%	11%
Student comments on *conclusions* to drawn from the study.	I think this was bad evidence because it did not relate to the model or talk about the nucleus. This is why it's bad evidence.	17%	14%	11%
Student identifies a feature of the evidence presentation	It has examples, pictures, and labels.	11%	64%	36%
Student refers to an aspect of the methodological processes				
Observation/measurement issues	… he actually looked through a microscope to get to his conclusion.	56%	47%	0%
Inference issues	… the picture was bad, and there are a lot of inferences based on her observation on the hamster.	11%	8%	6%
Design issues (e.g., controls)	I think evidence #2 is good evidence because the experiment was controlled.	14%	11%	17%
Execution and procedural issues	1. They [went] step by step. I really liked it. 2. …the images were unclear and fuzzy.	3%	0%	0%
Sample size issues	… their sample size wasn't so great. They only used one cat to put jellyfish gene in.	8%	0%	29%
Researcher credibility	… he said he hoped he didn't do anything wrong so he could have messed up.	14%	3%	11%
Corroboration from others	Furthermore he observes the flagellum and approves of it with other classmates to have all of the info covered.	0%	17%	0%

Results

A brief summary of several key results is as follows:

First, students' overall evaluations of all three studies roughly reflected the quality of the evidence as intended by the design team. The hamster evidence (intended to be the weakest evidence) was evaluated positively overall by 31% of the students, whereas the stronger flagella evidence and cat evidence were rated positively by 67% and 62%, respectively.

Second, in their responses, students frequently evaluated studies on the basis of the studies' conclusions or the results, rather than focusing on the methodological processes of the study. For example, one student wrote: "It is OK evidence because it does sound like a low number of mitochondria," focusing on the results of the study rather than the methods. These are not appropriate responses, in our view. Scientists, too, may view a study as good in part because it has a clear-cut, compelling result that informs theory (Staley, 2004).

Third, most students (over 75%) gave fairly detailed, level 3 responses to the questions. Thus, students were able not only to give responses; they gave reasoned responses providing justifications for their ideas.

Fourth, students collectively provided a broad range of comments focused on strengths and weaknesses of the methodological processes used. Table 1 displays the responses by students in selected categories for each piece of evidence. Collectively, students successfully identified most weaknesses and strengths of each study.

Fifth, students' responses exhibited some sensitivity to the particular strengths and weaknesses of the three pieces of evidence. Although space precludes a detailed analysis in support of this point, the different percentages of responses in each category across the three pieces of evidence supports the claim that students were sensitive to particular feature of each piece of evidence.

Sixth, a large majority of students' responses were normatively appropriate. That is, a majority of students' identified methodological strengths can indeed be regarded as strengths, and their identified weaknesses can be regarded as weaknesses.

Discussion

The two studies provide evidence that seventh graders are collectively aware of important criteria for evaluating the quality of evidence. Study 1 shows that it is possible for students in class discussions to collaboratively develop accurate evaluations of evidence quality, pooling the insights of different students. Study 2 shows that students collectively have a wide range of productive insights about evidence evaluation that can be so pooled. In other work (Authors, 2010, 2011b), we have found that when students collectively identify a broad range of important criteria for evaluating models, they develop class norms that successfully combine their productive individual understanding. Thus, these results provide a warrant for designing inquiry environments that encourage seventh graders to share their ideas about evaluating evidence so as to promote greater skill.

References

Allchin, D. (2011). Evaluating knowledge of the nature of (whole) science. Science Education, 95, 1-25.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic reasoning in schools: A theoretical framework for evaluating inquiry tasks. Science Education, 86, 175-218.

Chinn, C. A., Duncan, R. G., Dianovsky, M., & Rinehart, R. (2013). Promoting conceptual change through inquiry. In S. Vosniadou (Ed.) *International Handbook of Conceptual Change* (2nd ed.) (pp. 539-559). New York: Routledge.

Masnick, A., & Klahr, D. (2003). Error matters: An initial exploration of elementary school children's understanding of experimental error. Journal of Cognition and Development, 4, 67-98.

Reiser, B. J., Smith, B. K., Tabak, I., Steinmuller, F., Sandoval, W. A., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S. M. Carver & D. Klahr (Eds.), Cognition and instruction: Twenty-five years of progress (pp. 263-305). Mahwah, NJ: Erlbaum.

Rinehart, R. W., Duncan, R. G., & Chinn, C. A. (in press). A scaffolding suite to support evidence-based modeling and argumentation. *Science Scope*.

Staley, K. W. (2004). The evidence for the top quark: Objectivity and bias in collaborative experimentation. Cambridge: Cambridge University Press.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1008634. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.