PROMOTING THE DEVELOPMENT OF EPISTEMIC COGNITION

By

WILLIAM J. PLUTA

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

CLARK A. CHINN

and approved by

______________________________

______________________________

______________________________

______________________________

New Brunswick, New Jersey

May, 2015
ABSTRACT OF THE DISSERTATION

Promoting the Development of Epistemic Cognition

by WILLIAM J. PLUTA

Dissertation Director:

CLARK A. CHINN

Recent recommendations for reforming science instruction have advocated placing inquiry activities such as evidence evaluation, model-building, explanation, and argumentation at the center of the curriculum. An increasing number of studies indicate that instruction focused on these inquiry activities promote deep content learning and an informed understanding of the nature of science. A major challenge in any inquiry program is developing methods to move students’ reasoning forward. I report on the results of three studies aimed at exploring strategies to make science inquiry instruction more productive. The first two studies focus on the development of students’ ideas about epistemic criteria. Study 1 examines 324 middle school students’ preinstructional ideas about epistemic criteria for good scientific models. Study 2 examines four class discussions in which students propose, vet, and adopt class lists of epistemic criteria. Collectively, these two studies provide important information about students’ preinstructional ideas about scientific models and epistemic criteria as well as strategies for promoting student understanding of epistemic criteria. The third study focuses on students’ ability to coordinate evidence in order to develop a more inclusive, accurate model as well as their understanding of epistemic criteria for good evidence. Study 3
describes the results of an interview study in which 29 middle school students attempted to coordinate multiple pieces of conflicting evidence.

Overall, the three studies provide insights into ways to design inquiry learning environments. In particular, they have implications for scaffolding inquiry--I argue that productive scientific discourse and inquiry activities can be scaffolded through reflection on epistemic criteria for evaluating the quality of models and evidence. Criteria provide important information on the extent to which student need to be introduced to science-specific criteria and the extent to which students need to be scaffolded when engaging with multiple pieces of evidence. Finally all three studies provide important theoretical information on students’ epistemic resources and the overall sophistication of their understanding of the nature of science.
# Table of Contents

Abstract of the Dissertation ............................................................ ii

Table of Contents ........................................................................... iv

List of Tables ................................................................................ vi

Introduction ................................................................................... 1

References ..................................................................................... 8

Study 1. Learners’ Epistemic Criteria for Good Scientific Models .......... 14

Abstract ....................................................................................... 15

Introduction ................................................................................... 16

Methods ....................................................................................... 32

Results and Discussion ................................................................. 38

Conclusions .................................................................................. 63

Summary ....................................................................................... 71

References ..................................................................................... 73

Study 2. Developing Shared Epistemic Criteria .................................. 83

Abstract ....................................................................................... 84

Introduction ................................................................................... 85

Method ......................................................................................... 98

Results .......................................................................................... 102

Discussion ..................................................................................... 125
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>129</td>
</tr>
<tr>
<td>References</td>
<td>130</td>
</tr>
<tr>
<td>Study 3. Coordinating Evidence: Learners’ Adaptive Strategy Use</td>
<td>138</td>
</tr>
<tr>
<td>Abstract</td>
<td>139</td>
</tr>
<tr>
<td>Introduction</td>
<td>140</td>
</tr>
<tr>
<td>Overview and Goals of the Study</td>
<td>161</td>
</tr>
<tr>
<td>Methods</td>
<td>162</td>
</tr>
<tr>
<td>Results</td>
<td>166</td>
</tr>
<tr>
<td>Discussion</td>
<td>198</td>
</tr>
<tr>
<td>Summary</td>
<td>205</td>
</tr>
<tr>
<td>References</td>
<td>206</td>
</tr>
<tr>
<td>Themes Revisited</td>
<td>214</td>
</tr>
</tbody>
</table>
List of Tables

Table 1.1. Criteria for Good Scientific Models: Student Responses……… 40
Table 3.1. Summary of Problems………………………………………………..163
Table 3.2. Interview Questions……………………………………………………165
Table 3.3. Strategies and Brief Description………………………………………169
Table 3.4. Count Data Across Problem Types for Question A………………..173
Table 3.5. Count Data Across Problem Types for Question B…………………174
Table 3.6. Responses to Question D………………………………………………179
Table 3.7. Responses to Questions F……………………………………………...186
Table 3.8. Models Proposed by Students for Frog Deformity Problem……191
Introduction

Recent recommendations for reforming science instruction have advocated placing inquiry activities such as evidence evaluation, model-building, explanation, and argumentation at the center of the curriculum (National Research Council, 2007; NGSS Lead States, 2013). A growing number of studies indicate that instruction focused on these reasoning activities promote deep content learning (e.g. Chinn & Malhotra, 2002; Raghavan, Satoris, & Glaser, 1998; Ruiz-Primo, Li, Tsai, & Schneider, 2010; Smith, MacLin, Grosslight, & Davis, 1998; Snir, Smith, & Raz, 2003; White & Frederiksen, 1998) and an informed understanding of the nature of science (e.g. Bell & Linn, 2001; Hogan, & Maglienti, 2001; Metz, 2011; Penner, Giles, Lehrer, & Schauble, 1997; Pluta, Chinn, & Duncan, 2011; Sandoval, 2003; Schwarz & White, 2005; Schwarz et al., 2009; Smith, Maclin, Houghton, & Hennessey, 2000). A major challenge in any inquiry program is developing methods to move students’ reasoning forward (e.g. White & Frederiksen, 1998; Zimmerman, Raghavan, & Sartoris, 2003).

I report on the results of three studies that explore strategies to make science inquiry instruction more productive. The first two studies focus on the development of students’ ideas about epistemic criteria for good scientific models. The third study focuses on the development of evidence evaluation skills. All three reports are unified by a focus on understanding the development of an informed understanding of the nature of science. A principle commonality is that all three studies can be viewed as providing insights into students’ understanding and application of epistemic criteria: Studies 1 and 2
explicitly examine students’ criteria for evaluating scientific models and Study 3 addresses students’ tacit ideas about epistemic criteria for evaluating scientific evidence.

I now set the stage for the three studies by highlighting core issues around the emphasis on epistemic criteria and coordinating evidence.

**Epistemic Criteria**

Epistemic criteria are the standards used to evaluate scientific products (e.g., models, evidence, arguments). For example, scientists use criteria for judging the quality of scientific models and guiding the choice between alternative models (Kuhn, 1977; Newton-Smith, 1981; Sober, 1988); some of these criteria include:

- Good models have high levels of conceptual coherence and clarity.
- Good models are compatible with theories in other scientific disciplines.
- Good models are appropriately parsimonious.
- Good models are consistent with empirical evidence.

An example of an epistemic criterion for evaluating evidence is to prefer robust, readily replicable evidence that has large effect sizes (Staley, 2004). Arguments that explicitly or tacitly refer to these and other epistemic criteria are common in the written and oral discourse of scientists (cf. Bazerman, 1988; Staley, 2004). Despite the importance of criteria in everyday scientific practice, there has been minimal research addressing science students’ tacit or explicit knowledge, understanding, and use of epistemic criteria (Chinn, Buckland, & Samarapungavan, 2011).

There are several reasons for an increased focus on epistemic criteria in science classes. First, the use of epistemic criteria is a significant scientific practice in its own right. Understanding criteria and criteria-related practices is an important part of learning
how to participate in science, as well as understanding the nature of science (NOS). The use of criteria is embedded in the modeling, argumentation, and evidence evaluation practices that are central to inquiry curricula. Second, a focus on criteria provides a fruitful instructional target for scaffolding. Making epistemic criteria explicit in classrooms is a good way to make the cognitive practices of science visible. Students can, for example, be encouraged to develop and revise their own criteria for model goodness, argument strength, and so on. Reflection on criteria can help students gain mastery of new and challenging inquiry practices. Finally, the social processes of debating, vetting, adopting, and applying epistemic criteria in group and class discussions can contribute to the development of powerful knowledge-building communities within classrooms (cf. Brown & Campione, 1996; Scardamalia & Bereiter, 2006). Public discussion about criteria can enable classes to develop norms that support better group and individual learning.

I report on three studies examining students’ understanding of epistemic criteria. In the first two studies, I focus on criteria for model goodness because of the growing emphasis on modeling in both the history and philosophy of science (e.g., Giere, 2004; Godfrey-Smith, 2006; Longino, 2002; Nersessian, 2002; Suarez, 2003) and in science education (e.g., Chinn, Duschl, Duncan, Buckland, & Pluta, 2008; Raghavan & Glaser, 1995; Schwarz et al., 2009; Stewart, Hafner, Johnson, & Finkel, 1992; White, 1993; Windschitl, Thompson, & Braaten, 2008). I draw on the philosophy of science to define models as idealizations that scientists use to represent aspects of the world for specified purposes (Giere, 2004, p. 742). The third study concentrates on students’ tacit
criteria for good evidence and is well aligned with the norm that good models are consistent with high-quality evidence.

The initial two studies focus on a series of instructional activities aimed at introducing students to explanatory models: Students worked in pairs to judge whether artifacts were explanatory models or some other type of model (e.g. a toy model, scale model, data model). Students also evaluated the quality of pairs of scientific models. The artifacts were presented on a handout created as part of the research project. No instruction on models or modeling was provided. Students individually recorded their ideas about criteria for evaluating scientific models. As a class, students collaboratively discussed their ideas about criteria for evaluating scientific models drawing on their individual lists. Students vetted and adopted a class list of criteria. Teachers served as facilitators of the discussion. The class list of criteria served as a tool for shaping model-based instruction and classroom norms. Students were encouraged to hold each other accountable to class criteria.

This first study in this set describes the introductory activity in which students evaluate models and examines individual learners’ conceptions about epistemic criteria for good scientific models. Study 2 examines students’ ability to collaboratively propose, vet, and adopt a shared, class list of criteria for good scientific models. The second study describes the movement and development of knowledge from the individual to the collective sphere. Collectively, these two studies describe an initial set of instructional activities aimed at introducing students to scientific modeling and model-based inquiry. The third study, while having a focus on coordinating evidence, provides another context for students to engage with scientific criteria with minimal instructional guidance. These
three studies provide important insights into the extent to which instructors and instructional materials must introduce and make normative scientific criteria explicit to learners.

Future research will explore students’ (and teachers’) success at using the criteria to support scientific reasoning over the course of the school year.

**Coordinating Complex Evidence**

Increasingly, learners are being asked to evaluate, coordinate, and use complex data and evidence to generate scientific explanations, models, and arguments during instruction (National Research Council, 2007). For example, when learning about climate change, focusing exclusively on data related to global temperature increases is insufficient; students should be able to coordinate the details of a full range of evidence, such as evidence related to (a) sea level rise, (b) ocean temperature, (c) shrinking ice sheets, (d) declining artic sea ice, (e) extreme events, and (f) ocean acidification. Despite an emphasis of experimentum crucis in historical and popular accounts of science, scientists typically invest significant time considering multiple pieces of evidence from diverse sources. One important component of evidence coordination is applying epistemic criteria for both the evidence as well as the the models aligned with evidence. For example, students may need to draw on criteria for good evidence in order to reject or down grade the importance of particular studies.

Encouraging students’ to reason about multiple pieces of evidence supports the development of a sophisticated reasoning skill that scientists regularly engage in and promotes an understanding of criteria. Further, engaging in authentic reasoning (utilizing criteria) supports the development of an informed understanding of the nature of science.
Recent education policy debate has focused on 21st Century skills—coordinating multiple pieces of evidence may be representative of these types of skills (cf. Roschelle et al., 2011).

Researchers and instructional designers are increasingly developing instructional modules in which learners are presented with multiple pieces of evidence (e.g. Bell & Linn, 2000; Chinn, Duncan, Duschl, Buckland, & Pluta, 2008; McNeill, Lizotte, Krajik, & Marx, 2006; Sandoval & Reiser, 2004; Schwarz et al., 2009). A good example is Linn and colleagues WISE instructional environment (Bell & Linn, 2000). In one instructional module on light propagation, students were presented with 13 pieces of evidence that could be used to develop an explanation for light phenomena (Bell & Linn, 2000). In another WISE module, students critiqued six pieces of evidence bearing on whether rats are appropriate models for human in scientific research (Linn et al., 1999). While educational researchers have found that learning environments focused on evidence evaluation can support learning (e.g. Liu, Lee, Hofstetter, & Linn, 2008; Smith, Maclin, Grosslight, & Davis, 1998; White & Frederiksen, 1998), there is minimal research focusing on students’ ability to specifically coordinate the evidence itself. If students cannot consistently coordinate evidence, this may be a good target for instruction and/or scaffolding.

Overall, it is unlikely that learners can successfully coordinate the details of evidence. While there is little direct evidence bearing learners’ abilities to coordinate of multiple pieces of evidence, there are a number of studies examining other aspects of evidence evaluation. In general, this research suggests that without focused instruction or careful scaffolding, students will struggle at evidence-related tasks (cf. Zimmerman,
For example, students have difficulty (a) differentiating between models from evidence, (b) linking models and evidence, and (c) articulating explanations for how models fit the evidence (e.g. Bell & Linn, 2000; McNeill, Lizotte, Krajik, & Marx, 2006, Pluta, Buckland, Chinn, Duncan, & Duschl, 2008; Toth, Suthers, & Lesgold, 2002). In light of these results a number of instructional scaffolds and techniques have been designed to support these skills. It is likely that developing students’ evidence coordination would also require significant scaffolding—if not direct instruction.

Study 3 examines students’ ability to evaluate and coordinate evidence. It examines students’ ability to reconcile conflicting evidence. Two important findings of Study 3 are that (a) students struggled to attend to details of studies during evidence coordination, and (b) students’ have a wide range of normative and nonnormative criteria for evaluating scientific methodology. This study provides important information on how to scaffold authentic scientific inquiry.

**Summary**

I present three studies that provide important insights into how to improve student reasoning. These studies provide information on how to support the development of sophisticated epistemic cognition and reasoning.

Overall, all three studies indicate that students have sophisticated epistemic resources. For example, students have rich ideas about (a) the nature of models and modeling (Studies 1 and 2), (b) the role and interpretations of evidence in science (Study 3), and (c) socio-epistemic norms (Studies 2 and 3). In fact, the results of these three studies indicate that students’ epistemic cognition is significantly more sophisticated than previous research has suggested. One explanation for why these results conflict with the
conclusions of other researchers is that material that students completed in my studies included significantly more contextual information for students to draw on. I will emphasize the importance of providing students’ with appropriate contextual information when assessing epistemic cognition in Study 3. These results have important implications for theories related to epistemic cognition and for assessing epistemic cognition.

These results also indicate that science students may need more targeted instruction on specific epistemic forms—target structures which guide inquiry (Collins & Ferguson, 1993). For example, in Study 3 we found that few students demonstrated an understanding/or ability to use epistemic forms related to scientific interaction and causal chains. However, students do have important ideas or resources related to evaluating evidence.

References


Study 1. Learners’ Epistemic Criteria for Good Scientific Models

This chapter was published in the Journal of Research in Science Teaching in collaboration with Clark A. Chinn and Ravit Golan Duncan:

Abstract

Epistemic criteria are the standards used to evaluate scientific products (e.g., models, evidence, arguments). In this study, we analyzed epistemic criteria for good models generated by 324 middle-school students. After evaluating a range of scientific models, but before extensive instruction or experience with model-based reasoning practices, students generated lists of criteria of good scientific models. Students’ individual lists of criteria were compared to expert criteria, identified by philosophers of science, and with findings from previous research on students’ understanding of modeling. The most commonly listed criteria referred to the clarity, pictorial form, and explanatory function of models. Almost a quarter of the students included criteria relating to model fit with evidence. Students’ criteria provided insights into their understanding of the explanatory and descriptive goals of models; the constitutive, communicative, and epistemic features of models; and the role of evidence in supporting models. Collectively, students demonstrated familiarity with a wide range of modeling ideas that can be leveraged in instruction to promote deeper understandings of the modeling practice. We argue that inquiry-based science instruction should include a strong emphasis on epistemic criteria.
Introduction

Epistemic criteria play an important role in science; they are the standards scientists use to evaluate the validity and accuracy of scientific products such as models, arguments, and evidence (e.g., Kuhn, 1977; Laudan et al., 1986; Longino, 2002). For example, scientists use criteria for judging the quality of scientific models and guiding the choice between alternative models (Kuhn, 1977; Newton-Smith, 1981; Sober, 1988); some of these criteria include:

- Good models have high levels of conceptual coherence and clarity.
- Good models are compatible with theories in other fields.
- Good models are appropriately parsimonious.
- Good models are consistent with empirical evidence.
- Good models have a history of making novel empirical predictions.

An example of an epistemic criterion for evaluating evidence is to prefer robust, readily replicable evidence that has large effect sizes (Staley, 2004). Arguments that explicitly or tacitly refer to these and other epistemic criteria are common in the written and oral discourse of scientists (see, e.g., Bazerman, 1988; Staley, 2004).

Despite the importance of criteria in everyday scientific practice, there has been minimal research addressing science students’ tacit or explicit knowledge, understanding, and use of epistemic criteria (Chinn, Buckland, & Samarapungavan, 2011). The purpose of this study is to address this gap by examining middle-school students’ explicit ideas about epistemic (and nonepistemic) criteria for model goodness.

We focused on criteria for model goodness because of the recent emphasis on modeling in both the history and philosophy of science and in science education. In the
history and philosophy of science, recent work has emphasized the central roles that the
development, testing, evaluation, and refinement of models play in scientific practice
(e.g., Giere, 2004; Godfrey-Smith, 2006; Longino, 2002; Nersessian, 2002; Suarez,
2003). In science education, model-based instruction has been a focus of many research
programs (e.g., Chinn, Duschl, Duncan, Buckland, & Pluta, 2008; Raghavan & Glaser,
1995; Schwarz, Reiser, Davis, Kenyon, Acher et al., 2009; Stewart, Hafner, Johnson &
Finkel, 1992; White, 1993; Windschitl, Thompson, & Braaten, 2008). As a result of this
research, recent recommendations for reforming instruction have advocated placing
modeling activities at the center of the curriculum (National Research Council, 2007). A
major challenge in any inquiry program, including model-based programs, is developing
methods to move students’ reasoning forward. An important dimension of growth in
reasoning is students’ understanding and use of epistemic criteria. Therefore,
understanding students’ ideas about criteria for model goodness is an important step in
developing better methods for promoting reasoning.

There are several reasons for an increased focus on epistemic criteria in science
classes. First, the use of epistemic criteria is a significant scientific practice in its own
right. Understanding criteria and criteria-related practices is an important part of learning
how to participate in science, as well as understanding the nature of science (NOS). The
use of criteria is embedded in the modeling, argumentation, and evidence evaluation
practices that are central to inquiry curricula. Second, a focus on criteria provides a
fruitful instructional target for scaffolding. Making epistemic criteria explicit in
classrooms is a good way to make the cognitive practices of science visible. Students can,
for example, be encouraged to develop and revise their own criteria for model goodness,
argument strength, and so on. Reflection on criteria can help students gain mastery of new and challenging inquiry practices. Finally, the social processes of debating, vetting, adopting, and applying epistemic criteria in group and class discussions can contribute to the development of powerful knowledge-building communities within classrooms (cf. Brown & Campione, 1996; Scardamalia & Bereiter, 2006). Public discussion about criteria can enable classes to develop norms that support better group and individual learning.

In this article, we report on a study of middle-school students’ naïve (preinstructional) ideas about epistemic criteria for model goodness. We report individual students’ lists of criteria generated early in the school year, before the students had extensive experience with modeling and before they had discussed epistemic criteria as a class. Students evaluated models before generating criteria in order to provide a sense of what is meant by “scientific model” and to activate preexisting knowledge about scientific models. This research addresses the following questions: (a) What are middle school students’ initial ideas about modeling and the quality of models before they engage in intensive modeling practice or extensive work with criteria? (b) To what extent do student-generated epistemic criteria match the criteria used by practicing scientists? (c) What does the overall sophistication and diversity of students’ initial criteria suggest about appropriate and effective instructional approaches and strategies that can promote growth in understanding of scientific models and modeling criteria? Our analyses provide information about students’ understanding of modeling and science, leading to important insights into how to develop instruction that can help students advance beyond their naïve ideas about epistemic criteria.
Before describing the study, we review relevant literatures, including (a) models and epistemic criteria in science, (b) model-based instruction and epistemic criteria, (c) research on students’ use and understanding of epistemic criteria, and (d) students’ epistemic criteria and their conceptions of models. We conclude with a discussion of how an understanding of students’ naïve epistemic criteria can provide insights into the development of better inquiry-oriented instruction.

**Models and Epistemic Criteria in Science**

Like many science education researchers (Hogan & Maglienti, 2001; Penner et al., 1997; Windschitl et al., 2008; see also National Research Council, 2007) we draw on philosophers’ work in conceptualizing scientific models. Models are idealizations that scientists use to represent aspects of the world for specified purposes (Giere, 2004, p. 742). Scientific hypotheses make claims about similarities between models and real systems (Giere, 1988, p. 81), claims that can be empirically tested (Longino, 2002; Solomon, 2001). Well-known models include representations of evolutionary mechanisms, including genetic drift, natural selection, and transmutation; models representing the orbits of planets, such as the Copernican heliocentric and Ptolemaic geocentric models; and models depicting atoms such as the plum-pudding and Bohr’s models. Much of the work of scientists centers on developing models and investigating the extent to which models resemble the world in the intended respects (Giere, 1988; Kitcher, 1993; Nersessian, 2002; Niiniluoto, 2002).

Some of the model-specific cognitive practices performed by scientists include developing and revising explanatory models in response to evidence, evaluating the internal and external consistency of models, choosing among models, conducting
experiments and other studies to test models, and evaluating the quality and strength of evidence. Social activities are essential to carrying out these cognitive practices; these include distributing and coordinating the work of experimentation and observation, communicating findings within the scientific community, and engaging in argumentation and co-construction of knowledge. Discourse is central; scientists critique each others’ models, provide support and elaboration to further develop models, and appropriately reconcile alternate models, approaches, and ideas. Many of these strategies and practices are incorporated into model-based learning environments.

The use of epistemic criteria or standards is central to many of the practices outlined above. To decide whether a model is adequate, or whether one model is better than another, scientists and science students alike must evaluate models utilizing standards or criteria of model goodness. As we noted earlier, these criteria can include conceptual coherence, evidential fit (e.g., scope of evidence covered, degree of fit with that evidence), and parsimony. Similarly, when judging the supportive strength or fit between models and evidence, reasoners must evaluate the strength of evidence. Criteria for evaluating evidence strength include standards used to judge the methodological strengths and weaknesses of empirical studies. For instance, scientists may maintain that the strongest evidence derives from empirical studies that have appropriate controls, adequate sample size, use appropriate measures, and include multiple measures that give converging results. Epistemic criteria may be deployed with or without explicit awareness.

Criteria related to model-evidence fit and coherence with other well-established theories can be viewed as central to evaluating a scientific model. These primary
epistemic criteria center on the likely accuracy of the model, by which we mean the extent to which the model resembles intended aspects of the world in desired respects, as indicated by fit with empirical observations. Other criteria—such as criteria for how clearly models are presented—are secondary criteria because they do not directly impact the accuracy of a model; an accurate but poorly presented model can be represented in a clear way without changing any substantive content. In the present study, we classify student-generated criteria in terms of whether they are primary epistemic criteria, which are central to the epistemic practices of science (identified by philosophers of science) because they focus on the accuracy of the model, or secondary epistemic criteria, which do not directly impact the accuracy of the model, but contribute to epistemic aims of science, such as communicating important ideas clearly so that others can use these ideas. Distinguishing between primary and secondary criteria provides a useful guide for evaluating students’ epistemic cognition and development during generation and use of criteria for good models. It provides a rough indicator of the extent to which students are focused on accuracy of the model versus other model criteria that are important, but can be met by inaccurate as well as accurate models. Instruction aimed at introducing or developing understanding and use of criteria can then be targeted at specific levels of epistemic criteria.

Model-Based Instruction and Epistemic Criteria for Good Models

There is promising evidence that model-based instruction improves students’ ability to construct their own models (White, 1993), develop explanations and conclusions from models (White & Frederiksen, 1998; Schwarz & White, 2005), competently coordinate and work with multiple models (Gutwill, Frederiksen, & White;
1999), synchronize models and evidence (Pluta, Buckland, Chinn, Duschl, & Duncan, 2008; Snir, Smith, & Raz, 2003; Zimmerman, Raghavan, & Sartoris, 2003), critique and revise models (Penner, Giles, Lehrer, & Schauble, 1997; Stewart, Hafner, Johnson & Finkel, 1992), and engage in causal and analogical reasoning (Raghavan, Sartoris & Glaser, 1998; Harrison & Treagust, 2000). Model-based inquiry can promote conceptual change, as well (Chinn & Samarapungavan, 2009). Overall, these studies indicate that students engage in impressive reasoning when given the opportunity to practice modeling.

As model-based inquiry environments become more prevalent, there is a need for research that investigates methods of scaffolding reasoning in these environments. Existing research emphasizes helping students make critical structural distinctions, such as the distinction between explanations and evidence (e.g., Bell & Linn, 2000; McNeil, Lizotte, Krajcik & Marx, 2006; Suthers & Hundhausen, 2003, Toth, Suthers, & Lesgold, 2002). Other research explores scaffolds that support understanding of an inquiry cycle for modeling (e.g., Schwarz & White, 2005; Schwarz et al., 2009) and meta-awareness of core reasoning processes (White & Frederiksen, 1998). Scaffolding the development of epistemic criteria could be another very productive method of promoting growth in reasoning and understanding of the nature of science.

Scaffolds for epistemic criteria include methods that make student thinking about epistemic criteria public and encourage students to discuss criteria with peers and facilitators. One example of such a scaffold would be the public development of criteria for model goodness. Teachers could lead a class discussion in which students, having examined a variety of better and worse models, develop a list of public class criteria for
model quality. In further class discussions throughout the year, students could periodically revise the class criteria as they gain more experience with evaluating models. They would use their own criteria to evaluate their own, peer, and teacher provided models throughout the year, and the criteria could also be the basis of rubrics used by the teacher to evaluate students’ work. We have assayed this approach in our work with model-based inquiry in middle schools (Chinn, et al., 2008; Chinn, Pluta, Buckland, Rogat, DiFranco, & Witham, 2010).

In order to understand how to develop this and other scaffolds of epistemic criteria, it is valuable to understand students’ ideas about epistemic criteria. Instruction may profit from taking students’ naive ideas into account—much as knowing students’ prior conceptions about photosynthesis or about force and motion can guide the design of more effective instruction (Chinn & Brewer, 1993; Driver, Squire, & Wood-Robinson, 1994). At present, little is known about what students consider to be characteristics of good scientific models or how they think good models differ from poor models. With this information, we can develop instruction to help students learn progressively more sophisticated criteria.

Learning about epistemic criteria affords opportunities for students to learn about significant aspects of the nature of science. First, as students learn about epistemic criteria, they should learn that criteria cannot be applied rigidly or algorithmically; criteria must be applied contextually. For instance, it is often not straightforward to determine which of two or more alternative models has more supporting evidence, nor is it always obvious which of two theories is more parsimonious. Criteria can also conflict with each other (Laudan, 1977), as when a substantially more parsimonious model does
not explain the evidence quite as well; different scientists can reasonably come to
different conclusions about how to weigh competing criteria differently (Kuhn, 1977).
Second, students can learn that rival scientists sometimes differ in favored epistemic
criteria (Kuhn, 1977; Toulmin, 1958). For example, the history of biology has sometimes
seen divergence in the criteria of some scientists who give precedence to experimental
results and others who give greater significance to careful observations of nature (Mayr,
1982). This disagreement stems from a difference in criteria for what counts as the best
evidence. Third, students can learn that epistemic criteria can change. As an example,
consider the field of chemistry. Before Lavoisier, one core criterion for chemical
explanations was that good models must explain color and texture, and changes in both.
This criterion was much less central after Lavoisier. Over time, scientists critique the
prevalent criteria; they may abandon or modify old criteria and develop new ones
(Laudan, 1984; Longino, 2002). Like theories and models, criteria are accepted through
surviving repeated rounds of social criticism. Criteria that do not survive this criticism
may be abandoned, modified, or deemphasized.

Research on Students’ Use and Understanding of Epistemic Criteria

There is an increasing amount of research on students’ ideas about criteria. Five
prominent examples are studies by Samarapungavan (1992), Hogan and Maglienti
(2001), Penner et al. (1997), Smith, Maclin, Houghton, and Hennessey (2000), and
that first, third, and fifth graders made theory choices that conformed to normative
epistemic criteria recommended by philosophers of science, including range of
explanation, non-ad hocness of explanation, and empirical and logical consistency. A
majority of students could explicitly justify their theory choice utilizing the criteria of explanatory scope (e.g., by saying that one theory explains more than the other) and of empirical consistency (e.g., by saying that one theory isn’t as good because some evidence goes against it). However, a majority of students did not explicitly articulate the normative criteria to justify correct choices that favored simpler theories or theories that avoided ad hoc explanations. In this study, students articulated reasons that drew on criteria, but they were not asked to reflect metacognitively or talk about the criteria themselves. Moreover, only those few criteria that were relevant to the forced theory choices were examined in this study; whether students would spontaneously propose these criteria (or other criteria) was not investigated.

Hogan & Maglienti (2001) compared students’, laypeople’s, and scientists’ reasoning about scientific conclusions, finding that students and laypeople appealed to personal inferences and values as a criterion for judging the validity of the conclusions, whereas scientists did so far less frequently. Scientists also appealed to criteria relating to precision and specificity in the conclusions; students did not. Students were not asked to describe or reflect on the criteria they used to evaluate conclusions.

Penner et al. (1997) found that first and second graders who had engaged in modeling instruction were more likely to evaluate models by appealing to criteria related to functional aspects of models, whereas students who had not engaged in modeling instruction were more likely to draw on criteria related to perceptual qualities of models. The researchers did not ask students to be explicit about their general criteria. Similarly, Smith and colleagues (Smith et al. 2000) found that middle school students who engaged in constructivist science curriculum that included modeling generated more criteria for
evaluating scientific beliefs than students who participated in more traditional science instruction.

In an instructional study, Schwarz and White (2005) had students evaluate models using four criteria that were directly taught to them: accuracy, plausible mechanism, consistency, and utility of models. Students evaluated their own, peer-generated, and researcher-generated models utilizing the provided criteria. Schwarz and White found that students who used model-criteria demonstrated a better understanding of the nature of modeling, inquiry, and physics content than students who completed the same instructional unit, but without explicit use of criteria (comparison data were drawn from a different study described in White & Frederiksen, 1998). This result provides evidence that instruction focused on epistemic criteria can scaffold science learning. However, this study does not provide information on whether students have the ability to generate and use their own criteria, or what students’ own ideas about criteria are.

Although these studies provide valuable information about students’ epistemic criteria for good models, none of them employed a method that seems promising in understanding students’ epistemic criteria—namely, asking students to reflect on their own criteria for what makes a model good. The study described in this article addresses this gap in prior research by asking students to report on their ideas about criteria met by good scientific models.

**Students’ Epistemic Criteria and Conceptions of Models**

Students’ conceptions about epistemic criteria are likely to be interconnected with their understanding of models. For example, if students hold the conception that models are literal copies of nature, they will likely fail to understand why models needs to be
revised in light of evidence and, more specifically, will be unable to apply criteria related to the nuances of model-evidence fit. Similarly, if students fail to understand the idea that models frequently present mechanisms, then criteria such as “has a good explanatory mechanism” will be meaningless to them. This raises the question of whether students have a good enough understanding of models to even begin considering criteria for good models.

There have been only a few studies of students’ (or teachers’) understanding of models and modeling (Grosslight, Unger, Jay, & Smith, 1991; Treagust, Chittleborough & Mamiala, 2002; Schwarz & White, 2005; Windschitl et al., 2008). Most research suggests that students’ views of models and modeling are inaccurate without instruction (e.g., Carey & Smith, 1993; Abd-El-Khalick, Bell & Lederman, 1992; Grosslight et al., 1991; Treagust et al, 2002). For example, Lederman and colleagues have found that few people recognize that scientists’ evaluations of models are tentative and somewhat subjective (Lederman et al., 2002). Grosslight et al. (1991) found that in the absence of instruction, most secondary-school students thought of models as little more than direct replications or copies of reality (67% of mixed-ability seventh graders; 23% honors-level eleventh graders). Few understood that models are created for specific purposes (12% of seventh graders; 36% of eleventh graders), understood the role of evidence in developing models (0% of seventh graders; 45% of eleventh graders), or realized that models can generate predictions (Grosslight et al., 1991). Treagust and his colleagues (2002) reported similar results, with two major differences. Treagust et al. (2002) found that over 70% of the students (ages 13 to 15) in their sample believed that models are revised in light of evidence, and nearly 50% agreed that models are used for “making predictions,
formulating theories and showing how information is used” (p. 365). They also found that a preponderance of students viewed models as direct copies of nature.

Given the interrelatedness of epistemic criteria and modeling, eliciting students’ epistemic criteria may provide more information about students’ understanding of modeling. Although the current study does not attempt to be an inclusive survey of students’ understanding of modeling, it may potentially shed light on the reasons for some discrepancies in previous research. In addition, research on students’ understanding of formal science does not provide much insight into the epistemic knowledge students actually use during instruction (Louca, Elby, Hammer, & Kagey, 2004; Sandoval, 2005). Student-generated lists of criteria may more closely align with the knowledge students actually use and display during learning and inquiry.

Overview and Goals of the Study

The present study investigates criteria that students explicitly generate after examining and evaluating the quality of a range of scientific models. The present study is the first to explore students’ own generated sets of epistemic criteria. Generating these lists required metacognitive reflection on the criteria themselves. Students were presented with multiple models of varying purposes and qualities, which they discussed in pairs. Then individual students generated their own lists of criteria for what made good models.

These data potentially provide important information about students’ epistemic cognition. If students can generate reasonable sets of criteria on their own, then it will be possible for students’ criteria to become the basis for classes to construct community norms to use epistemic and other criteria for model goodness. Teachers can lead discussions in which students propose and critique various criteria for model goodness,
culminating in an agreed-upon set of shared criteria. This would be a highly student-centered approach to developing shared norms regarding criteria. Before attempting to engage students in this type of discussion, a teacher would likely want to know the extent to which students collectively understand primary and secondary criteria and the extent to which these ideas are shared by students within a class.

For example, if the present study shows that students generate a combination of primary, secondary, and nonnormative criteria, the conversation could be geared towards the dialogic vetting of criteria for their effectiveness in choosing between alternate models; the teacher could perhaps focus discussion towards primary criteria or distinguishing between primary and other kinds of criteria. If students mostly generate primary epistemic criteria, then teachers could engage students in sharing their ideas, without needing to redirect the conversation toward criteria the students do not propose on their own. If, on the other hand, students cannot generate appropriate criteria, it suggests that more teacher-centered approaches may be needed to enable students to develop criteria. One of the aims of this study is to determine whether seventh-graders’ naïve understanding of criteria provides an adequate basis for a student-centered approach, or whether more teacher-centered approaches are needed.

In short, the central goal of this study was to identify students’ early, developing ideas about epistemic criteria for model quality. The primary source of data was student-generated lists of criteria, developed before students received extensive instruction on modeling, but after they had the opportunity to reflect on several scientific models similar to those found in textbooks. As mentioned previously, we addressed these questions: (a) What are middle school students’ initial ideas about modeling and the quality of models
before they engage in intensive modeling practice or extensive work with criteria? (b) To what extent do student-generated epistemic criteria match the criteria used by practicing scientists? (c) What does the overall sophistication and diversity of students’ criteria imply about appropriate and effective instructional approaches and strategies?

Prior research provides grounds for both pessimism and optimism about the quality of students’ generated criteria for good models. On the pessimistic side, there is a large body of research that appears to demonstrate that students have a poor understanding of the nature of science (Abd-El-Khalick et al., 1992; Carey & Smith, 1993), of models in particular (Grosslight et al., 1991; Treagust et al., 2002), and of theory-evidence relations (Zimmerman, 2007; cf. Bråten et al., 2011). These findings suggest that students’ criteria for model goodness might focus exclusively on issues of understandability or clear communication rather than on issues of evidential fit. This tendency could be reinforced by textbook models, which are intended to communicate scientific ideas rather than to be evaluated by students. Further, Schwarz and White (2005) found that students could use criteria that they were explicitly taught, but that most students could not reflectively articulate these criteria even after instruction. The ability to reflect on criteria requires a degree of metaawareness of one’s epistemic cognition that might be beyond the reach of many seventh graders.

On the other hand, there are some reasons for optimism. First, Samarapungavan (1992) found that even elementary-school students can specify reasons implicating scope of evidence and consistency with evidence as reasons to prefer one theory over another. By the middle school age, students may have further developed the capacity to reflect metacognitively on their criteria. Second, previous research on models has not provided
students with much experience with models before asking them questions about models; it could be that students do not understand the researchers to be referring to scientific models so much as to physical models, such as a model of a ship. If students have a better understanding of what the term *model* refers to, they may be better able to articulate criteria for model goodness. In this study, we guarded against misunderstanding of the task by exposing students to a range of scientific models and asking them to reflect on their quality before they developed their lists of criteria. We believe that this procedure provided students with a context for thinking about scientific models.

The study was part of a yearlong project involving seventh-grade life science students and teachers who were implementing a model-based inquiry program called PRACCIS (Promoting Reasoning and Conceptual Change in Science). This program of research has many features in common with other model-based inquiry programs (e.g., Raghavan, Sartoris & Glaser, 1995; Sandoval & Reiser, 2004; Schwarz & White, 2005; Schwarz et al., 2009; Stewart et al., 1992; White & Frederiksen, 1998). This curriculum and research project (Chinn et al., 2008) aims to develop instructional schemes and tools which teachers can easily embed within their own instructional materials as well as to develop instructional modules lasting 1 to 4 weeks that teachers can use flexibly within the constraints of their own state and district curricula. A central instructional scheme in this project is the *reasoning seminar*. During reasoning seminars, students engage in argumentation in which evidence is used to construct, revise, and evaluate explanatory models.

The assessment of students’ own epistemic criteria for good scientific models reported in this article took place near the beginning of the year, before students had
extensively engaged in modeling and model evaluation with criteria. Before developing their list of criteria, students reflected on a range of good and poor scientific models. This initial, orientating activity helped reduce confusion about what was meant by the term scientific model and provided an opportunity to reflect on the characteristics of good models. Thus, the assessment which followed this activity served as a measure of students’ initial ideas about models and criteria for good models.

Methods

We examined lists of criteria generated by seventh grade students following a forty- to fifty-minute activity aimed towards activating students’ existing knowledge of scientific models. In analyzing the results, we compare and contrast the criteria identified by students with the criteria identified by philosophers of science and with findings of previous research on students’ understanding of modeling.

Participants

Participants were 324 students in four seventh-grade teachers’ classes in two diverse, suburban New Jersey school districts (two teachers per district). The two districts reflected different levels of success on statewide tests of proficiency in mathematics and reading. In School 1, 1% of students qualified for free or reduced-price lunch; 97% of students were white. Students performed well on New Jersey state exams; approximately 90% of students in the school reached proficient or advanced proficient levels on state language arts and mathematics exams. In School 2, 27% of students qualified for free or reduced-price lunch; 47% of students were black, 27% were white, 15% were Hispanic, and 11% were Asian. The school’s performance on New Jersey state standardized tests was lower than in School 1. Approximately 70% of students reached proficient or advanced proficient levels on the language arts exam; approximately 65% reached these
levels on the mathematics exam. Students in fifteen classes (at least three classes per teacher) participated in the study; each class was composed of students of mixed abilities.

Context

The task of generating individual criteria was embedded into each teacher’s regular instructional schedule. The assessment took place within the first month of school, immediately after parental consent and student assent were obtained. Students had not received significant instruction or experience with scientific modeling before the assessment. However, because teachers had received training and had adopted the model-based curriculum during the previous school year, some teachers used the word model or mentioned some modeling ideas to students before this activity. All four teachers had spent some time providing instruction on the scientific method (this instruction took place before the research team began collecting data; the research team did not facilitate any instruction during this period). This instruction was generally traditional instruction on topics such as measurement and experiments and did not focus on the core ideas that the modeling curriculum addressed—developing explanatory models to fit evidence, revising models in accordance with new evidence, incorporating mechanisms into models, considering the strength of evidence, understanding the nature of models, and so on.

Materials

Over a period of 40 to 50 minutes spanning all of one class period and part of a second, students completed a series of model evaluation tasks. We specifically aimed to: (a) provide a brief, working definition of models to students; (b) activate students’ existing knowledge of models, developed through viewing models in textbooks and other
media; and (c) present contrasting cases that would help students begin reflecting on specific features of models related to quality.

Students were provided with a 13-page packet. The packet included the following explication of scientific models: “Scientific models are explanations. We can use scientific models to explain things in the world. Often, we can also use scientific models to predict things. For example, a model of how plants grow explains why and how plants grow, and it can help us make predictions about when plants will grow and when they will not grow.” We had two reasons for presenting this definition of models: (a) Developing explanatory and predictive models was a primary instructional goal for the larger research project. (b) Explanatory models, which can also be used for predictions, are a prevalent form of model in contemporary science; hence, learning to develop, revise, select, and use explanatory models is an authentic practice of scientists. Although there are other types of scientific models, philosophers of science (e.g. Giere, 1996; Kitcher, 1993; Longino, 2002; Machamer, Darden & Craver, 2000) and science educators (e.g. Clement, 2000; Duschl, 2008; Schwarz et al., 2009; Stewart, 1992; Windschitl et al. 2008) have emphasized the centrality of explanatory models in science. Researchers such as Schwarz (Schwarz et al. 2009), Smith (Smith et al. 2000), and Windschitl (Windschitl et al. 2008) have focused on developing instructional schemes that promote students’ understanding and use of the explanatory models. Other researchers have examined data models (e.g. Lehrer & Schauble, 2004), emergent models (e.g. Penner, 2000), or focused on the analogical or cognitive aspects of models and modeling (e.g. Harrison & Treagust, 2000). These are important types of models in science, as well, but we restrict our focus to explanatory models in this article.
The first section of the packet displayed twelve different representations of volcanoes. Students circled the representations they thought were models and discussed their ideas with a partner. The representations included models, non-models, and representations that were debatable as to whether they were models. Non-models included pictures of actual volcanoes erupting and a volcano toy. We included both descriptive models (e.g., a static diagram of a volcano with only physical features labeled) and explanatory models. Four different model representations explaining how volcanoes erupt were included (flowchart, written explanation, causal diagram, and pictorial model with text). We also included a model explaining how dome volcanoes collapse, a formula for the force of a volcanic eruption, a data representation (a map displaying volcanic sulfur dioxide concentrations), and a schematic diagram of scale model. This activity aimed to provide grounds for reflecting on what kinds of representations are scientific models and what kinds are not scientific models, and to encourage students to think about scientific models they had previously viewed. Most students have seen scientific models in textbooks, television, or other media; however, they likely have not interacted with them in an active way (e.g. evaluating, choosing between alternative models, or revising them). Nor were students likely to have ever considered instances that are not models and thought about what distinguishes between models and non-models.

The second section of the packet presented seven pairs of models, each pair on a single page. The models addressed phenomena with which most students were somewhat familiar, including butterflies/lifecycles, global warming, food webs, amphibian lifecycles, the water cycle, plant growth, and diffusion. The two models in each model
pair modeled the same general phenomena using two different representations. The dimensions along which the models differed (which correspond to the seven phenomena) included, (a) whether the models were descriptive or explanatory, (b) whether the models were explanatory or representations of data, (c) the degree of model complexity, (d) the presence and type of communicative features (e.g., labels, title), (e) the extent to which the models included details, including both relevant and extraneous details, (f) whether mechanisms were present and, if so, the types of mechanisms, and (g) the extent to which models were consistent with everyday data that students were likely to know about.

Each pair of models had one of two types of questions for students to consider. The first type of question was generic, “Which model is better? Or are they equally good? Or is it impossible to say which one is better?” The second type of question referred directly to a purpose for the model, such as, “Which of these two models is better if you want to explain how the smell of perfume can spread across a room?” These contrasting cases and the questions were designed to help students reflect on specific features of models that were related to model quality. We expected students to be relatively familiar with the phenomena modeled in this section and likely to be familiar with some of the specific representations.

**Procedures**

At the beginning of class, teachers informed students that they would be learning about scientific models. Teachers briefly discussed the explanation of models found on the first page of the packet. Students then worked through the introduction to modeling packet. For each page of the packet, students individually studied the representations and selected the best models. Students then discussed each page with a partner. As teachers
monitored student progress, they encouraged students to explain their thinking to each
other but generally sought to avoid providing specific information or feedback to
students. Upon completing the packet, students individually generated and wrote six
criteria for good models, following these instructions: “You have been thinking about the
characteristics of good models and not-so-good models. Now, individually, make a list of
the most important characteristics of good (rather than bad) models. Write down six
important characteristics of good models that you can think of below.” In most classes,
students completed the introduction to model packet in one day and generated their
individual criteria on a second day. By the time students generated their individual
criteria they were likely to have understood what we were referring to when we used the
term scientific models. Students’ individual lists of criteria generated after the orientating
activity are the sole focus of the analyses reported in this article.

Coding

We iteratively refined and developed codes through several analyses. In
developing our coding scheme, we initially developed candidate categories based on
criteria developed by philosophers, criteria suggested by the literature on students’
understanding of models (especially Grosslight et al., 1991, and Treagust et al., 2002),
and an analysis of our introduction materials. Candidate criteria developed by
philosophers included categories such as range of evidence explained, consistency with
evidence (and avoiding inconsistency), testability, coherence, and simplicity. Candidate
criteria suggested by the work on modeling included reference to kinds of models (e.g.,
explanatory and descriptive), use of different representational forms, and clarity. Our
initial coding scheme included 47 categories. We revised the coding scheme by
repeatedly examining student responses, adding to and modifying the initial candidate categories to better fit student responses. The initial categories were winnowed to 32 categories. Student responses assigned to each category were iteratively checked against each other to make sure that similar responses were consistently being coded in the same way.

Students often wrote the characteristics of good models as short phrases, such as “[a good model is] neat, detailed, descriptive” or “[a good model] shows how or why.” In such cases, we judged that the words in these phrases reflected different criteria. For example, in the first example, we coded the word neat as an instance of the criterion “good models are well organized,” which was defined as including neatness. The word detailed was classified as an instance of the category “good models are detailed.” And the word descriptive was coded as an instance of the criterion “good models are descriptive.” The second example, “shows how or why,” was coded as an instance of explanation criterion because an explanation has as a primary function giving an account of how or why things occur (Kitcher, 1981). All categories that emerged from the coding are described in the results section.

We coded all written lists of student-generated criteria. Intercoder agreement, computed based on 50% of the data coded by two coders, was 85%. Disagreements were resolved through discussion.

**Results and Discussion**

On average, each student generated 6.1 distinct criteria, and each class collectively generated a mean of 25.5 distinct criteria. Table 1.1 presents the percent of individuals providing each criterion, as well as the percent of classes in which at least one student generated each criterion. Overall, there were no major differences at the class.
teacher, or school level in the distribution of individual criteria. We organized the criteria into five broad categories, which we discuss below. First, we discuss students’ criteria
<table>
<thead>
<tr>
<th>Criteria Elements</th>
<th>Definition and Example</th>
<th>Students (n =324)</th>
<th>Classes (n =15)</th>
<th>Interpretation of Epistemic Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goals of Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explains-1</td>
<td>Explicitly uses the word <em>explain</em> or <em>explanation</em>, e.g., the model should explain something.</td>
<td>51%</td>
<td>100%</td>
<td>Primary. Philosophers frequently discuss the explanatory role of models (e.g. Giere, 1988; Machamer et al., 2000; Mayr, 1982). Because the introduction-to-models activity explicitly used the word <em>explains</em>, we distinguished between responses that used this word and responses that used other words.</td>
</tr>
<tr>
<td>Explains-2</td>
<td>Uses other descriptive language for explanation, e.g., shows how and why.</td>
<td>4%</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>Describes</td>
<td>Explicitly uses the word <em>describe</em>, e.g., a good model uses <em>descriptive</em> pictures.</td>
<td>24%</td>
<td>87%</td>
<td>Secondary. Although they do not address the central explanatory goals of science, purely descriptive models have played a role in the history of science (e.g., anatomy, botany), as well as in educational settings (Kuhn, 1977; Mayr, 1982).</td>
</tr>
<tr>
<td>Provides information</td>
<td>Suggests models consist of static or isolated pieces of <em>information</em> or <em>facts</em>, e.g., provides information.</td>
<td>41%</td>
<td>100%</td>
<td>Vague. Although models do inform, this is not a central feature of models as discussed by philosophers. We think this language is too imprecise to take as evidence of a deep understanding of modeling.</td>
</tr>
<tr>
<td>Answer a question</td>
<td>Alludes to the idea that models provide answers to problems; responses usually use the words <em>answer</em> or <em>question</em>.</td>
<td>21%</td>
<td>73%</td>
<td>Vague. It is unclear what students mean by this response. Different students might have different kinds of questions in mind.</td>
</tr>
<tr>
<td>Examples</td>
<td>Models are examples (models as illustrations of explanations), but not necessarily as abstract exemplars (an explanations), e.g., gives an example</td>
<td>8%</td>
<td>53%</td>
<td>Vague. This language is too imprecise to suggest a well-developed understanding of models. It is unclear whether students are referring to a concrete example or a more idealized example.</td>
</tr>
<tr>
<td>Data</td>
<td>Explicitly uses the word <em>data</em>, without suggesting that the data support the model.</td>
<td>5%</td>
<td>53%</td>
<td>Secondary. Models of data are an important kind of model in science (Chinn &amp; Brewer, 2001; Staley, 2004). Students may be suggesting that models can present data or that models are built upon or supported by data.</td>
</tr>
</tbody>
</table>
### Model Constituents

<table>
<thead>
<tr>
<th>Component</th>
<th>Definition</th>
<th>Primary %</th>
<th>Secondary %</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pictures</td>
<td>Explicitly uses the word <em>pictures</em></td>
<td>60%</td>
<td>100%</td>
<td>Secondary. Responses in the Model Constituents category specify features of models that are needed to help communicate what the models represent. Many views of science emphasize the importance of disseminating information to scientific peers (e.g., Bishop &amp; Trout, 2005; Goldman, 1999). If scientists cannot understand models, even if the models meet many other primary epistemic criteria, the models are valueless. Thus, although these criteria are not central criteria, they do have significant secondary value.</td>
</tr>
<tr>
<td>Words</td>
<td>References the use of words, texts, and language in models.</td>
<td>30%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Diagrams</td>
<td>Explicitly uses the word <em>diagrams</em></td>
<td>19%</td>
<td>93%</td>
<td></td>
</tr>
<tr>
<td>Labels</td>
<td>Explicitly uses the word <em>label</em></td>
<td>13%</td>
<td>93%</td>
<td></td>
</tr>
<tr>
<td>Arrows</td>
<td>Explicitly uses the word <em>arrows</em></td>
<td>12%</td>
<td>87%</td>
<td></td>
</tr>
<tr>
<td>Visuals</td>
<td>Explicitly uses the word <em>visuals</em></td>
<td>11%</td>
<td>73%</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Explicitly uses the word <em>title</em>.</td>
<td>7%</td>
<td>63%</td>
<td></td>
</tr>
</tbody>
</table>

### Communicative Elements

<table>
<thead>
<tr>
<th>Component</th>
<th>Definition</th>
<th>Primary %</th>
<th>Secondary %</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarity</td>
<td>Appeals to lucidity, e.g., model is <em>easy to understand</em>; model is <em>clear</em>.</td>
<td>64%</td>
<td>100%</td>
<td>Secondary. As noted above, many views of science emphasize the importance of disseminating information to scientific peers. For these reasons, these criteria are conductive to the spread of scientific knowledge, and we classify them as secondary criteria.</td>
</tr>
<tr>
<td>Focus</td>
<td>Good models stay on topic.</td>
<td>37%</td>
<td>93%</td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>Refers to models being neat and organized. Arrows <em>not all over the place</em>.</td>
<td>26%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Audience</td>
<td>References the audience that the models are designed for. Includes peers, scientists, or general public.</td>
<td>5%</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>Details</td>
<td>Appropriate details are represented in good models, e.g., just enough details.</td>
<td>40%</td>
<td>100%</td>
<td>Primary. Ideas about appropriate details and complexity appear closely related to the normative criterion of parsimony (Kuhn, 1977; Popper, 1959). In addition, philosophers have discussed how models connect to the world in appropriate respects and to appropriate degrees; models with appropriate detail and complexity meet these purposes (Giere, 2002; Longino, 2001).</td>
</tr>
<tr>
<td>Complexity</td>
<td>Appeals to the appropriate complexity represented in good models, e.g., not too complex.</td>
<td>10%</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>Sequence</td>
<td>Models have sequence, e.g., has steps, it has a timeline, it has a cycle.</td>
<td>29%</td>
<td>100%</td>
<td>Secondary. Ideas about sequence and steps invoke ideas about causation, mechanism, and explanation, all key elements of scientific models (Machamer et al., 2000).</td>
</tr>
</tbody>
</table>
However, because students’ responses stopped short of referring to mechanisms, we assign it secondary rather than primary significance.

### Evidential Criteria

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Explicitly refers to <em>evidence</em> supporting the model, using the word <em>evidence.</em></th>
<th>19%</th>
<th>93%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other support</td>
<td>Refers to evidence supporting the model without using the word <em>evidence,</em> e.g., Data supports the model</td>
<td>8%</td>
<td>60%</td>
</tr>
<tr>
<td>Quality of support</td>
<td>Refers to the quality of evidence or reasons, e.g., the evidence <em>should be true</em></td>
<td>6%</td>
<td>73%</td>
</tr>
<tr>
<td>Quantity of support</td>
<td>References the quantity of evidence or reasons, e.g., <em>needs several pieces of evidence.</em></td>
<td>2%</td>
<td>40%</td>
</tr>
<tr>
<td>Reasoning</td>
<td>Explicitly refers to reasons, e.g., gives reasoning.</td>
<td>9%</td>
<td>87%</td>
</tr>
</tbody>
</table>

### Epistemic Elements

<table>
<thead>
<tr>
<th>Quantity of explanation, description, information</th>
<th>A model has a significant amount of explanation/description/information expressed in the model, e.g., it has a lot of explanation, explains lots of stuff.</th>
<th>20%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Appeals to an idea concerning how faithfully the model depicts the target, e.g., the model is correct.</td>
<td>8%</td>
<td>80%</td>
</tr>
<tr>
<td>Interest</td>
<td>Suggests that models should be</td>
<td>5%</td>
<td>53%</td>
</tr>
</tbody>
</table>

Primary. The centrality of evidential criteria in model evaluation has been emphasized by many philosophers. There are a number of specific sub-criteria related to evidential support, including both the quality of the evidence that are used to support a theory as well as the quantity of evidence that provides support.

Secondary. Although reasoning is central in science, students’ criteria concerning reasoning were generally unelaborated.

Primary. Explanatory scope is an important criterion discussed by philosophers of science (Kuhn, 1977; Laudan et al., 1986). A model with explanatory scope describes a wide range of data or observations, and explains data the model wasn’t designed or intended to explain.

Primary. Kuhn (1977) and others have discussed *Accuracy* as a key criterion for model evaluation. Other philosophers often discuss accuracy in terms of empirical success (Goldman, 1999; Solomon, 2001).

Primary. Criteria related to *Interest* captures a sense of
interesting or creative.

Realism

Explicitly use the words *real* or *realistic.*

2% 33%

Secondary. Philosophers have debated whether scientific models should be viewed in a realist, or at least a truthlike, sense (e.g., Niiniluoto, 2003). Students who give this criterion have, like many scientists, adopted a realist stance. However, exactly what students mean by “real” or “realistic” is unclear; they could mean that it is accurate, or that it has real entities, or simply that it is plausible. Hence, due to its relative lack of clarity, we have assigned this code to the secondary category.

Quality

A model has a high quality explanation/description/information expressed in the model, e.g., a good model has a great explanation.

19% 93%

Vague. Task instructions requested that students make a list of the most important characteristics of good (not bad) models. Students’ responses which imprecisely noted model goodness may have been simply repeating the task instructions.
related to the goals of models, because understanding the goals of modeling is essential to understanding why other modeling criteria are important. Second, we discuss model constituents because these responses provide information about students’ basic ideas about what models should “look like” and thus how models are structured to accomplish their goals. We then consider communicative criteria, evidential criteria, and other epistemic elements in an order that roughly corresponds to the overall prevalence of responses within each category.

Within these categories, we further distinguished between three levels of criteria: (a) primary criteria, which are central to the practices of science (as observed by philosophers of science) and center on the likely accuracy of the model (by the terms accuracy, we refer to the extent to which the model is congruent with the world in desired respects, as indicated by congruence with empirical observations); (b) secondary criteria, which do not directly impact the accuracy of the model, but contribute to epistemic aims of science; and (c) criteria that are vague or suggest misconceptions about the practices of science.

In the following sections, we discuss individual and class criteria within each of the five categories of responses. In each section, we first present results and then discuss implications of our results. We also consider the implications for teachers who wish to lead class discussions in which classes propose and discuss criteria that can then serve as community norms for the evaluation of models; knowledge of the distribution of individual criteria in a class can influence how teachers go about facilitating these discussions.

Goals of Models
Criteria classified in the Goals of Models category reflect students’ ideas about the purposes and functions of models in scientific practice. As we noted earlier, a central aim of contemporary scientific research is to develop explanatory models—models that explicate mechanisms and identify casual and functional relationships. However, educational research suggests that students often view models as having the goal of literally depicting objects (Grosslight et al., 1991; Treagust et al., 2002). Students’ criteria relating to the goals of models have the potential to provide information on whether students see models as carbon-copy replicas of phenomena or whether they appreciate the broader goals of models.

In our sample, 90% of the students made some reference to the goals of models in their criteria. Thirty-seven percent of students noted only a single model goal, 33% noted two model goals, and 20% of students specified three or more criteria related to model goals on their individual criteria lists. Thus, 53% of the students in the study generated multiple goals for models in their lists of criteria. Individual students proposed an average of 1.5 criteria in this category. In the four most common responses, students wrote that models should provide explanations (55% of students), information (41% of students), descriptions (24% of students), and answers to a question (21% of students).

The explanation category suggests an understanding of explanation as a central goal of modeling. Because philosophers have identified explanation as an important goal of models (e.g., Giere, 1988; Kitcher, 1993), we classified this criterion as a primary epistemic criterion. In our coding scheme, we distinguished between two categories indicating that the purpose of models is to explain. The Explains-I code (51% of all students) refers to the explicit use of the word explain (or morphologically derivative
words such as explanation). The Explains-2 code (4% of all students) refers to other language used by the student that relates to explanation (e.g., a good model “shows how and why”). We differentiated between these two kinds of responses because our instructions at the beginning of the two-lesson sequence of activities explicitly mentioned that models provide explanations. Hence, although many students whose responses were coded as Explains-1 might have understood that models explain phenomena, others may have just been repeating what they had been told (but note that these students would have had to recall this information more than 24 hours after encountering it).

We characterized two types of goals (providing descriptions and modeling data) as secondary epistemic criteria. Forty-one percent of all students wrote that models provide descriptions. In the sciences, descriptive models play an important role, particularly in the applied sciences, and often serve as a foundation for the development of explanatory models. Descriptive models are also ubiquitous in science education materials (e.g., models identifying part of the cell, models identifying the location of planets), thus students may be very familiar with this goal of models. But description is less central to modern science than is explanation; as we noted above, in most fields scientists seek to develop explanatory frameworks. A smaller number of students suggested that models are data (5%)—as opposed to stating that models are supported by data. Data models are indeed a very important and specific kind of scientific model (Chinn & Brewer, 2001; Lehrer & Schauble, 2006; Staley, 2004), but they are distinct from explanatory models. Thus, we deemed ideas relating to models as data as a secondary level criterion for the quality of the explanatory models that students had been exposed to. It is unclear whether students providing this response had an understanding
that data can be modeled or whether they were confused about the distinction between models and data supporting models.

Three responses suggested a vague or nonnormative understanding of models. Twenty-one percent of students indicated that models provide answers to questions, but the kinds of questions that students were referring to was unclear. We suspect that if these students had been prompted to elaborate or provide examples, the questions generated would be a mix of explanatory, descriptive, and narrow information-seeking questions. Responses in the information category (41% of all students) suggested that students believed models were providing static or isolated pieces of information. Of the students who provided responses that fell within the information category, 80% proposed other primary and secondary criteria that suggested a deeper understanding of the goals of models (106 students). The third response suggesting a vague or nonnormative understanding of models was to view models as themselves being examples (8% of all students); this response seems compatible with a view of models as illustrations of explanations rather than as the actual proposed explanations. However, only 7% of the students who generated responses in the examples category (2 students) failed to generate an additional type of model goal. Thus, most students who gave this response had a view of models that was not limited to models as examples.

In considering the entire corpus of criteria relating to model goals, students in three classes collectively proposed all six criteria, eight classes proposed five different criteria, three classes proposed four different criteria, and one class proposed three criteria. All classes had multiple students who advanced explanations and information as criteria. Thirteen of 15 classes had multiple students who advanced description as a goal.
Thus, teachers who wish to draw on students’ ideas about the goals of models within a class discussion would likely find that a wide variety of model goals, including the most important primary and secondary goals, would be generated by one or more students in any given class.

In summary, a majority of students demonstrated some understanding of multiple model goals, and students in all classes collectively noted a wide range of goals. The diversity of student responses surprised us, given that previous research on students’ understanding of modeling has reported that most students see models only as direct replications of reality (Grosslight, 1991; Treagust, 2002). Judging from the number of students who suggested that models are examples, we suspect that some students see replica-type models as an important kind of model. However, this was a small percentage of students (only 8%). Our orientating activity may have steered students away from a misinterpretation of the word model as simply referring to the everyday usage of the word as a scale replica. The orientating activity presented students with models with a variety of different goals; the majority of these models were explanatory, but we also included descriptive models, replica models (which we do not count as scientific models), and even models of data.

Although students were introduced to the idea of scientific models with a definition that stated that models are explanations, we were still surprised by the number of students that referred to explanation in their criteria. One possible account for this finding is that students simply remembered and repeated what they were told about models at the beginning of the orientating activity on the previous day. However, our introductory definition also emphasized that models can be used to make predictions, and
no students mentioned predictions or any related idea in their criteria. If students were simply repeating the introductory definition, they would have also mentioned ideas about predictions in their criteria. Thus, students either already knew something of the explanatory role of scientific models, or the idea of explanations fit so readily with their prior knowledge that they were immediately able to adopt this idea with minimal instruction.

Class level data show that all classes collectively put forward primary epistemic ideas. Thus, teachers who elect to lead class discussions about criteria can expect students to raise significant goal-related criteria, without having to suggest these criteria themselves.

Providing students with more opportunities to evaluate models with their classmates may encourage students to develop an appreciation for the epistemic value of explanatory models, as well as develop more nuanced criteria regarding purposes, such as “explanatory models are often more useful than descriptive models,” or a criterion that suggests that “the kind of model should match its purpose.” Criteria relating to the goals of models may be particularly easy for students to evaluate, as suggested by the high number of students who proposed criteria within this category. These questions should be explored in inquiry curricula.

Model Constituents

Scientific models take a variety of representational forms, including pictorial diagrams highlighting mechanisms, mathematical relationships, and even explanatory text. The constituent parts of these models—including color, titles, diagrams, and captions—aid in the important communicative function that models play when
scientists present their work. Responses in the Model Constituents category specify features of models that help communicate what the models represent.

Eight criteria fell within this category. According to the three most common responses, students believed that models need pictures (60% of all students), words (30%), and diagrams (19%). Students also mentioned labels (13% of all students), arrows (12%), and titles (7%). Individual students mentioned an average of 1.5 model feature criteria. Twelve classes proposed six or more of these criteria; two classes proposed five; one class proposed four criteria within this category.

The connection between these constitutive parts of models and the epistemic aims of science is indirect. Without these parts, models would not exist; yet not all diagrams with arrows and titles are scientific models. For this reason, we think these criteria are best categorized as secondary epistemic criteria.

Taken as a whole, these criteria suggest that most students view models as something other than a literal or scale copy of an everyday object. A literal copy (e.g., a toy car being a model of a real car), does not require titles, arrows, and labels. But models that describe or explain complex phenomena would require these communicative tools to be properly represented.

As we have noted, in the orientating activity students evaluated whether the different representations of volcanoes and volcanic processes were or were not models; these representations varied in whether they had some of the constitutive features mentioned by the students. Students also viewed pairs of contrasting scientific models that included pictures, words, diagrams, and arrows. One specific example of a model students viewed and discussed was a picture representing the life cycle of a frog that did
not include a title, label, or captions. Thus, we see a link between the orientating activity and students’ criteria, as students’ lists included constituents that were missing from this model.

Model constituent criteria are likely very accessible to most students. We think that these ideas might be a good starting point for introducing model critique activities to young students. As students become more adept at identifying and using these model features, they can progress to considering more central epistemic criteria. In addition, this category of responses reflects familiarity with the kinds of models students see in traditional instruction (e.g., pictorial or diagrammatic models in textbooks). As students are introduced to different kinds of models (e.g., mathematical models and explanatory models emphasizing unobserved mechanisms), new types of model features should emerge in criteria lists, and students may also realize that few, if any, features appear in all types of models. For instance, many students in our sample believed that good models require pictures; however, as students gain familiarity and skill at working with more abstract representations, they may come to believe that pictorial representations are not necessary, and in some cases can contain unnecessary and distracting details.

Communicative Elements

Criteria in the *Communicative Elements* category are closely related to those in the Model Constituents category. The Communicative Elements criteria reflect more general ideas about how models are typically designed in order to communicate ideas appropriately and clearly. The focus is not on the constituents of the models *per se* (which also play a communicative role) but in how the constituents are organized to make the models understandable to others. The three most common responses in this category
concerned *appropriate clarity* (64% of students), *appropriate details* (40%), and the *focus of model representations* (37%). Responses classified as indicating a “focus of model representations” suggested that the model is designed specifically for the topic at hand. For example, if a model is supposed to depict photosynthesis, it should not instead depict transpiration.

Each of the seven criteria in this category was proposed at a relatively high frequency, with the average individual proposing 2.1 criteria involving communicative elements. Students in eight classes collectively proposed all seven criteria in this category. Four classes proposed six of the criteria, and three classes proposed five of the criteria, with complexity and audience being the only two criteria missing in these classes. Thus, teachers who wish to lead a discussion about criteria following individual generation of criteria can expect most or all of the communicative criteria in Table 1 to be mentioned by one or more students in each class.

Ideas about clarity, organization, focus, and audience are primarily communicative and apply to any kind of scientific model. Models cannot be evaluated for accuracy or fit with evidence if those who peruse them cannot make out what they mean; thus, models must meet criteria of successful communication before they can be evaluated for epistemic quality. Because a false or even fraudulent model can successfully meet communicative criteria, we judge that most communicative criteria are secondary rather than primary epistemic criteria. As social epistemologists such as Bishop and Trout (2005) have noted, the successful spread of knowledge requires effective communication of ideas. A brilliant, empirically successful model that is
presented in a way that other scientists cannot understand will not gain acceptance. Thus, communicative criteria are very important at the secondary level of epistemic practice.

Communicative criteria, especially criteria concerning clarity, organization, and focus are likely well established within school cultures. These criteria would also apply to tasks such as writing an essay in language arts or completing a project in history class. It is somewhat surprising that more students did not suggest these criteria. One possibility for why more students did propose these criteria is that many students considered them too obvious or not specific enough to the task of creating a list of criteria for good models.

Two criteria in this category, “models should have appropriate details” and “models should be appropriately complex,” seem to relate directly to ideas about parsimony. Parsimony is a primary epistemic criterion discussed in the philosophy of science literature on theory and model choice (e.g. Kuhn, 1977). A priori justifications of parsimony generally are centered on the idea that adherence to the parsimony criterion is one defensible way to work around the problem that models are underdetermined by supporting data. Naturalized approaches to epistemology point to a number of important episodes in science in which scientists explicitly adhered to this criterion (e.g., development of evolutionary explanations; Sober, 1981). We note, however, that although students proposed criteria for the value of appropriate details and complexity, they did not explicitly mention the value of idealizations. Actual scientific models are idealizations that capture only those aspects of a phenomenon that are of interest to the modeler. Thus, they do not contain pictorial details that are irrelevant to the model purposes (e.g., a model of photosynthesis processes in a microbiology journal does not
include a picture of the flowering parts of the plants, which are irrelevant to photosynthesis). When students mentioned appropriate details as criteria, it is not entirely clear whether they were advocating leaving out details that are hard for the audience to understand or details that are unimportant for the purposes of the model.

Students also frequently mentioned the importance of sequence in models (29% of all students). *Sequence* evokes ideas related to causation. Developmental research has shown that even very young children can invoke causal mechanism when describing how objects work (e.g., Wellman & Gelman, 1992). Temporal and visual contiguity, as would be found in a pictorial or written sequence, have been found to encourage reasoners to invoke causal explanations (White, 1988). This provides more support that many, if not most, students recognized the explanatory role of scientific models. We have categorized the sequence criterion as a secondary epistemic criterion because there is only an indirect link to causation and explanation. As students gain more experience with causal and explanatory models, we expect that students will begin to articulate the conditions for good causal mechanisms in criteria.

A number of the criteria within the Communicative Elements category appear to relate directly to the contrasting models that students viewed during the orienting activity. For example, students compared a complicated food web showing the relationships among thirteen species with a simple food web that only showed three species. These contrasting models might have spurred ideas about appropriate levels of simplicity or complexity. Similarly, students compared a water cycle model that presented only relevant, well-defined structures and mechanisms, with one that included additional, unnecessary and cute pictorial details serving only an aesthetic function (i.e., cartoon
images of rain, lakes, clouds, and a landscape which included houses). These models may have spurred students’ ideas about appropriate detail. These models also all had a strong sequential component, which could have encouraged students to articulate sequence as a criterion.

Overall, the high frequency and relative sophistication of the proposed communicative criteria suggest that, collectively, students begin with ideas that provide a foundation for developing more nuanced criteria. For example, students’ ideas about model complexity and detail can be the basis for discussions about the appropriate degree of parsimony. Similarly, instruction could aim to move students’ vague ideas about sequence to more explicit criteria describing what a good mechanism entails. Thus, students’ rich preinstructional ideas provide resources for developing more nuanced understandings. Similarly, there are a number of well-established “school” criteria that students proposed, such as clarity. It may be appropriate to discuss these criteria when introducing model criteria, but because most students are familiar with them, teachers may want to emphasize these criteria less than criteria that are less familiar to students. Explicit meaning-making discussion should likely be centered on primary and secondary criteria that are less established in students’ repertoires.

**Evidential Criteria**

The category *evidential criteria* comprises responses that referenced evidence and reasoning. Evidence is central to the enterprise of model-based science (Giere, 1988; Longino, 1990, 2002; Niiniluoto, 2002). The epistemic justification for a model is ideally, and most commonly, based on the quantity and quality of evidence supporting the model. Further, a model can be questioned if there is evidence that contradicts the
model. Thus, responses within this category reflect awareness of model goodness criteria that are central to the epistemic practices of science.

Overall, 24% of students directly indicated that good models are supported by evidence. Most of these students (19% of all students) used the word evidence; 8% of all students used other language that suggested understanding of evidential support. Three percent of students wrote one criterion in each of these two categories. Nine percent of students noted that models should include reasons or reasoning. A few students mentioned the importance of the quality and quantity of evidence or reasons (6% and 2% of students, respectively). The average individual proposed 0.4 criterion related to evidence.

All but one class included at least one student who proposed a criterion that used the word evidence. The majority of classes had at least one student who used other language to suggest evidence, noted ideas about the quality of evidence, or mentioned reasoning. In 43% of classes students proposed criteria concerning the amount of evidence. Eighty-seven percent of classes had at least three students who proposed one of the criteria in the evidential criteria category.

The relatively high use of the words evidence and reasoning suggest that many of these middle school students had some awareness of the role of evidence in epistemic practices. This is a particularly interesting result because our orientating activity did not make any direct reference to the role of evidence in evaluating models and because research suggests that many students struggle to explicitly differentiate between evidence and theories or models (Zimmerman, 2007). Had we inserted some rudimentary evidence
or ideas about evidence into our orientating activity, perhaps an even larger number of students would have specified a role for evidence in evaluating models.

As discussed above, previous studies have provided conflicting results on students’ conceptions of modeling revolved around the role of evidence in modeling (Grosslight et al., 1991; Treagust, et al., 2002). The results of this study are congruent with Treagust et al.’s (2002) finding that students have some understanding of the role of evidence in model evaluation. Treagust et al. (2002) found that most students agreed with statements suggesting that models are to be revised in light of evidence. Our measure was not sensitive to whether students’ simply understood that evidence was somehow involved in modeling practice, that evidence is used to choose between models, or that models can actually be revised in light of evidence. Our study shows that many students spontaneously generate criteria related to evidence, without prompting.

Students’ criteria concerning reasons and reasoning were vague. Students likely have many ideas about what constitutes a reason. A reason could tout the logical coherence of a model, note a model incoherence, or it could specify how evidence supports a model. A reason could also be nonepistemic grounds for accepting or rejecting a model. For example, Hogan & Maglienti, 2001 found that students appealed to their own prior beliefs when evaluating conclusions, rather than drawing on provided evidence. Within the context of this task, we think that responses concerning reasons and reasoning suggest that some students have an emerging understanding that scientific models are not static artifacts, but are developed and used in concert with evaluation by scientists (or students within an inquiry environment).
As discussed previously, Kuhn (1977) emphasized the importance of criteria for the accuracy and explanatory scope of a model. Model quality is higher when there is a good fit between models and evidence. The presence of ideas about evidence in students’ lists of criteria indicates that many students are aware of normative criteria of accuracy with respect to evidence. A few of the middle-school students even displayed an even more nuanced view of evidence by appealing to the quality and quantity of evidence. There is no reason, of course, to think that students’ understanding of accuracy is as nuanced as philosophers or scientists. For example, there was no indication that they understood evidential fit in terms of the closeness of quantitative fit with the predictions of a mathematical model. Nor did they specifically identify covering a diversity of evidence types as an important criterion (e.g., the success of evolutionary theory in explaining phenomena in many different domains, from molecular biology to paleontology). Nonetheless, there is evidence that some students have developed some significant precursors of these more sophisticated ideas.

Epistemic Elements

The Epistemic Elements category consists of five generic responses relating to model quality. Twenty percent of students wrote about the quantity of explanations, descriptions, and information presented in models (e.g., a good model has lots of explanation), 8% of students proposed criteria for accuracy, 5% proposed that good models are interesting or creative, 2% proposed criteria referencing ideas about realism, and 19% of the students referred simply to model quality in a general way. On average students proposed 0.5 criteria in this category. Most classes had at least one student who proposed criteria concerning quantity (93% of classes), accuracy (80% of classes), and
interest/creativity (53% of classes). Only a third of the classes had students who proposed a criterion relating to model realism.

Students who mentioned model quality may have simply been echoing the task instructions to make a list of the most important characteristics of good (not bad) models. For this reason, we have classified criteria concerning general model quality as being too vague to be assigned to either primary or secondary epistemic status.

Students who proposed the four other criteria within this category appear to have demonstrated an emerging grasp of several normative epistemic criteria. For instance, students who proposed criteria relating to the quantity (e.g., a good model has lots of explanation) may have held the normative conception that good models can explain a large array of phenomena and observations (Kuhn, 1977). We viewed this criterion as a primary epistemic criterion, due to the centrality of far-reaching explanations in science. Another interpretation is that students believed that a model is better to the extent that students (or scientists) have invested a greater quantity of work or effort into constructing or working on a model. This interpretation is more consistent with existing school norms, but because no students used additional language to support this interpretation, it is not well supported by overall pattern of results.

Students in this study, much like philosophers of science, wrote that models should be accurate. It is possible that these students were thinking of a more literal mapping between the model and the phenomena (Grosslight et al., 1991; Treagust et al., 2002), but it is also possible that students were referring to evidential fit of the sort discussed in the previous section. It is possible that students were simply giving voice to well-established school norm that students generate a factually correct response that
matches the teacher’s expected answer, such as when answering an algebra question, responding to a history teachers’ question, or completing science work in a traditional classroom (Duschl, 1990). Despite this possibility, we classified accuracy as a primary epistemic criterion due to its close association with fit with the evidence, as well as its implied association with truth, which many philosophers have taken to be a condition of knowledge (Alston, 2005). Moreover, we think it is more likely that if students were focused on teacher-provided or sanctioned ideas, rather than empirically-supported ideas, they would be more likely to write that the model was “right” or that it was the “right answer.” This is a case where teachers (and researchers) should collect more detailed information on students’ conceptions of accuracy, as well as aim to move students from the well-entrenched school conception to a scientific conception.

The students who proposed that good models are interesting or creative expressed an idea that philosophers have discussed as very important to the functioning of science. A model that is interesting to a student is probably presenting new, unexpected information to them. It appears to us that this type of criterion captures a sense of novelty that is central to scientific progress, and it also reflects the importance of uncovering not just any truth but significant truths that are of interest to the scientific community (Haack, 2003; Kitcher, 2001). We thus count this as a central epistemic criterion.

Finally, students who wrote that good models are real or realistic, may be demonstrating a more simplistic notion of models being literal replications of phenomena, or they may be expressing an idea that models should be accurate in reflecting reality in some way. Philosophers (e.g., Cartwright, 1983; Kuhn, 1977; Giere, 2004; van Fraasen, 1980) have debated scientific realism, that is, the question of whether
theoretical entities and processes of science refer in some way to counterparts in the world. Many scientists have adopted critical realist stances in which models are viewed as truthlike or resembling the world in important respects (Giere, 2006; Niiniluoto, 2002). Thus, students’ ideas about the realism of models may not be that different from many scientists. However, given that students did not express exactly what they meant by this criterion, we count this as a secondary epistemic criterion.

While completing the orientating activity, students evaluated two pairs of models that potentially contributed to the incubation of ideas related to the criteria within this category. For example, they evaluated two models representing diffusion, one depicting Brownian motion and the second depicting a non-normative model in which molecules reproduced and multiplied as if they were living things. From this pair, students may have recognized that it was unlikely that both models could be accurate, although most were probably unaware of which model was closer to the scientifically accepted model. Similarly, the two food web models, described above, may also have contributed to the development of these criteria; students may have decided that the more complex model explained a larger part of the ecosystem, and thus was more accurate or perhaps “real.”

**Coordinating Responses Across Categories**

A question raised by our data is the extent to which the most sophisticated responses were made by the same individuals. Was it the case, for example, that a small number of sophisticated students proposed the majority of the primary criteria? Or were the more normative responses spread across a broader range of students?

Of the 31 criteria in our coding scheme, we categorized ten criteria as suggesting an understanding of primary epistemic aspects of science, 17 criteria as suggesting an
understanding of secondary epistemic aspects of modeling, and four which were vague or suggested a non-normative understanding of modeling. According to our scheme, a student with the highest level of epistemic understanding would propose primary epistemic criteria suggesting that scientific models (a) provide explanations (explains-1 or explains-2), (b) require evidence (generated at least one code in the evidential criteria category), (c) are interesting, (d) represent appropriate detail or complexity, (e) have significant scope (quantity code), and (f) are accurate. No student proposed criteria within all six of these categories, and only 2% of students proposed criteria that fell into five categories. Eight percent of students proposed criteria in four categories and 18% proposed criteria which fell into three of these main categories. Thus, all students in our sample had room for significant epistemic growth at the individual level. We did find that 16% of all students proposed both explanation and evidence, perhaps the most important of the primary epistemic criteria.

It is also of interest to examine the distribution of vague criteria. A vague set of personal criteria would not allow students to effectively evaluate peer or teacher-provided models. We found that 34% of student had no vague criteria on their list. Twenty percent of students had over a quarter of their criteria categorized as vague. Only five students (1.5%) had half or more of their criteria coded as vague or non-normative. Thus, most students had some understanding of primary and secondary criteria which were not vague.

Finally, we examined the distribution of responses across the five categories (goals of models, model constituents, communicative elements, evidential criteria, and other epistemic criteria. Overall, seven percent of students proposed criteria that fell into
all five categories of criteria. Thirty-four percent proposed criteria within four categories and 43% proposed criteria in three categories. Thus, most students suggested criteria that spanned a fairly broad range of epistemic categories.

Conclusions

Our goal in this analysis was to explore middle school students’ ideas about criteria for judging the quality of scientific models. We extended previous research by both focusing on students’ own ideas as they developed lists of epistemic criteria and by interpreting these results at a collective as well as an individual level. Students demonstrated familiarity with a wider range of modeling ideas than has been documented by previous studies. Although many students proposed established, school criteria, these were seldom the dominant type of criteria on students’ lists. At a class level, most classes had multiple students who proposed primary, normative criteria.

Student Generated Epistemic Criteria

Students’ criteria provided insights into their understanding of the goals of models, constituent parts, communicative features, the role of evidence in modeling, as well as other epistemic features.

A number of epistemic criteria correspond to those used by practicing scientists. Of five prominent scientific criteria identified by Thomas Kuhn (1977), students generated ideas which resembled three:

Accuracy. Eight percent of individuals and 80% of classes specifically noted accuracy as a criterion, while 24% of individuals and all fifteen classes proposed criteria relating to evidence, which also appears to be related to the accuracy of models.

Explanatory scope. Twenty percent of individuals and 100% of classes noted the importance of quantity of explanations/descriptions/information.
**Parsimony.** Over 40% of individuals and 100% of classes proposed that models should have appropriate details or complexity.

Students did not generate any criteria related to Kuhn’s fourth and fifth criteria: internal or external consistency, or fruitfulness. Students also proposed seven high-level communicative criteria (e.g., *models are clear, appropriately detailed, etc.*), which philosophers have argued play a central role in the uptake of scientific ideas (Bishop and Trout, 2005). Considering that the students had limited experience with scientific models and that they had not engaged in discipline-specific model-based inquiry, we conclude that the students displayed a relatively-sophisticated understanding of communicative criteria for good models.

There are, of course, a number of criteria identified by philosophers of science that students did not generate and that could be introduced to students over time. Some of these include criteria relating to observational nesting (that is, a more successful model explains data that competing models explain and successfully explains additional data, as well), the historical track record of empirical success, and the smoothness with which adjustments can be made in the face of explanatory failure (that is, a model is preferred if its explanatory failures can be corrected with minimal modifications) (Newton-Smith, 1981).

Our results differ from some of the previous research on students’ naïve understanding of both the nature of science and modeling practices. For example, Grosslight and colleagues (1991) found that none of the seventh graders in their study elaborated on the role of evidence in constructing models; in contrast, many participants in the present study spontaneously noted that evidence was important. Both Grosslight
and Treagust and their colleagues (Grosslight et al., 1991; Treagust et al., 2002) concluded that students primarily see models as direct replicas of phenomena rather than abstract or idealized explanations. Students in the present study produced criteria that suggested that many understood a model’s explanatory function, by explicitly noting that models are explanations and that models often include representational devices, such as arrows, titles and labels to explicate models. If students saw models only as scale, literal representations, they would have not seen these devices as necessary.

A conservative interpretation of our results is that students were merely able to identify appropriate vocabulary that could be used to generate model quality criteria. If this were the case, students’ understanding of modeling and epistemic criteria could be interpreted as superficial and unsophisticated (Grosslight et al., 1991). However, this interpretation is not supported by the overall pattern of results. As discussed previously, only one vocabulary word (explanations) appeared both in the orientating activity and in students’ criteria. Another important vocabulary word (predictions) appeared in the orientating activity (in the same paragraph with the word explanation), but no students generated criteria related to predictions. In addition, student used terms such as evidence, accuracy, complexity, and sequence, none of which appeared in materials that they had viewed. Therefore, a better-supported interpretation of the results is that students exhibited an understanding of a broad variety of criteria for model goodness, albeit with a lesser emphasis on criteria related to evidential support than on other criteria.

One possible reason for the better understanding of students in this study is that we provided them with more authentic examples of scientific models to consider as they generated their criteria for model quality. Grosslight et al. (1991) presented students with
a toy airplane, a map, a static picture of a house, and a textbook diagram of the water cycle early in their interview protocol. Treagust et al. (2002) did not provide participants with any examples of scientific models before they completed a questionnaire.

Participants in these two studies may have been confused about the kind of models they were evaluating; it seems likely that a toy model would be the first kind of model that comes to mind for a seventh grader, and even for many adults. Hence, participants may have assumed that toy models were the sort of thing that the researchers were asking about. In contrast, our task may have allowed students to grasp what sort of thing is meant by the term “scientific model,” rather than requiring them to guess. Although we think that a better understanding of the task improved students’ responses, it is unlikely that our participating students had had many substantive opportunities to engage with or evaluate scientific models. Thus, rather than drawing on any explicitly learned criteria for models, students were developing and crystallizing their criteria as they reflected on the models they were given to analyze.

Students’ responses appear to have been associated with the specific models they evaluated in the orientating activity. We think that this suggests one of two possibilities. One possibility is that seventh-grade students have existing ideas or resources for evaluating epistemic artifacts, honed during non-epistemic evaluative activities. Specifically, students may have general “evaluative resources” developed through everyday activities such as judging favorite music, fiction, or choosing a best or favorite superstar soccer player. Students have few opportunities to engage with these ideas during epistemic practices such as choosing between alternate scientific models; the model evaluation task may have prepared students to restructure their existing evaluative
resources around the epistemic aim of evaluating models (Louca et al., 2004). Recent
research has examined how different contexts, activities, and instructions may lead
students to approach activities in very different ways. For example, Engle, Nguyen, &
Mendelson (2012) concluded that transfer (of learning) was facilitated when instructors
adopted an “expansive frame,” characterized by encouraging students to think beyond the
specific learning task. Similarly, Scherr and Hammer (2009) interpreted particular student
behaviors and activities as being associated with multiple epistemological frames (e.g., an
opportunity for sense-making versus seeing an assignment only as busywork). The
orientating activity may have appropriately reframed the criteria generation task for
students such that it highlighted the types of knowledge and norms which were most
appropriate for the task.

The second possibility is that students constructed entirely new knowledge about
models and model evaluation through the orientating task and from interacting with their
peers. However, this conclusion is not well supported. We have only been able to
speculate about possible relationships between the models students viewed and the
criteria they generated; we saw no explicit links in their written criteria. In addition, if
students’ criteria were primarily derived from the model evaluation task, fewer students
would have proposed criteria relating to evidence and more students would have
generated criteria related to explanations. One limitation of the current study is that our
data are not robust enough to link students’ individual epistemic criteria directly to their
engagement with the models that they viewed. The orientating model-evaluation activity,
despite being relatively short, provided a context for thinking about models. We think
that students might have become more attuned to evidentiary considerations had the task
included more models that students could recognize as being consistent or inconsistent with familiar evidence. An interesting topic for future research is to investigate how variations in the exemplar models to which students are exposed affect the criteria that they generate. In addition, future research should examine criteria generated by students in different age bands, as well as criteria that are manifested during processes of learning and inquiry (cf. Sandoval, 2005; Greene & Azevedo, 2010, Greene, Muis, & Pieschl, 2010).

**Criteria and Instruction**

Given the rich array of criteria proposed by students, the results of this study provide important insights into the kinds of instruction that may improve students’ model-based reasoning and understanding of the nature of science. The task of generating epistemic criteria would likely provide teachers with information on their own students’ level of understanding with which they can make instructional decisions.

If students can generate reasonable sets of criteria on their own, then it would be possible for students’ criteria to become the basis for classes to construct community norms to use epistemic and other criteria for model goodness. In our view, the students in the current study did indeed generate a reasonable set of collective criteria. Although few individuals could generate most primary epistemic criteria on their own, nearly all primary epistemic criteria were evident in each of the fifteen classes in the present study. Therefore, the goal of instruction should perhaps focus on facilitating the sharing and vetting of these epistemic ideas. Social mechanisms may foster wide adoption of high-level epistemic criteria in the classroom. Anderson and colleagues (Anderson, Nguyen-Jahiel, McNurlen, Archodidou, Kim, Reznitskaya, et al., 2001) found that when
argumentation strategies are used in discussion, they tend to spread and occur with increasing frequency in classrooms with open-participation formats (as opposed to teacher-led discussion formats). The use and understanding of high-level criteria may spread and increase in a similar fashion. Student proponents of criteria could be encouraged to explain their rationale during discussion about criteria. Such discussions could help other students understand and appreciate the importance of specific criteria; further, students will be engaging in precisely the kind of dialogic process that scientists partake in when proposing and vetting criteria. This could develop nuanced understandings of normative epistemic standards, while also developing knowledge of broader community processes for establishing and vetting standards.

Centering inquiry on students’ own epistemic criteria may support the development of powerful classroom learning communities. Engle and Conant (2002) described a science classroom in which students were held accountable to disciplinary norms (e.g., sharing ideas and using evidence); observations suggested adherence to these norms led to greater authentic, disciplinary engagement. In general, the norms were established through teacher introduction, emphasis, highlighting, and encouragement. Having students generate their own epistemic standards, as we have done, may encourage the rapid transfer of this responsibility from teacher to students, by giving students a sense of ownership over criteria and assuring that at least some students understand the criteria.

The processes of sharing and using criteria could be scaffolded in a number of ways. First classes of students can work together to develop a class list of criteria. This class list could serve as the foundation for emerging epistemic norms in the classroom.
Students can revisit and revise this list as their ideas become more normative. This list would serve as a public record of the students’ collective, emerging understanding of criteria. Second, students and teachers can develop rubrics based upon criteria (drawn from the best or most common student ideas, from a class list of criteria, or even from the list described in this study). Students can then use the rubrics to evaluate both their own and provided models; this may lead to deeper content knowledge. Finally, having students rank criteria may help students identify key differences in epistemic levels (e.g., the differences between what we have dubbed as primary and secondary criteria). Future research should explore the efficacy of these approaches.

We think that epistemic criteria may serve as a tool with which to develop students’ understanding of nature of science (NOS). Most studies of students’ understanding of NOS suggest that students do not see theories or models as tentative and explanatory, nor do they grasp the dialogic processes of coordinating theories with evidence (Lederman et al., 2002). In this study, however, we have found that some students understand that fit with evidence is a criterion for models and that good models are explanatory. These incipient understandings of criteria for good models could be the basis for discussing the nature of science and encouraging a more sophisticated NOS understanding. Students’ initial criteria for good evidence and for good arguments could serve a similar role in discussions about the role of evidence and arguments in science.

As with most new instructional practices in science, engaging students in modeling in a sophisticated way is challenging because it is inconsistent with many well-established school, classroom and instructional norms (c.f. Chinn & Malhotra, 2002; National Research Council, 2007). One of the attractions of making students’ own criteria
the starting point for model-based instruction is that it provides students with a high degree of authority and ownership over the classroom norms (Blumenfeld, Kempler & Krajcik, 2006; Engle & Conant, 2002). Most inquiry programs attempt to foster student authority during the process of deciding on questions or developing intellectual products (e.g. models and experiments). Given that classroom norms shape and scaffold students’ choices when generating inquiry questions or developing products, epistemic criteria seem to be an advantageous starting point. This research suggests that students have the resources to develop a reasonable set of criteria. Future research should explore to what extent students use their own criteria and at what point students recognize the limitation of their criteria and amend them.

Summary

Science students are often thought to have unsophisticated knowledge of the epistemic practices of science. We found, to the contrary, that students seem to have a wide range of ideas about one important element of this practice—the epistemic criteria for good models. Many of the criteria proposed by students are similar to the criteria used by scientists (as identified by philosophers of science). Students collectively generated 31 distinct criteria for evaluating scientific models. Primary epistemic criteria included criteria related to the explanatory function of models, the role of evidence, appropriate details, and accuracy. Fifteen criteria were related to the communicative or constituent features of models; these suggested that students saw models as more than just direct copies or scale models. Students generated several different evidentiary criteria, although only about a quarter of individuals generated these criteria. Given the complexity of student responses, we think epistemic criteria are a potentially powerful starting point for instruction aimed towards improving student reasoning and understanding of the nature
of science. The distribution of student responses within classes suggest that instruction centered on peer sharing and vetting of ideas related to criteria would be an appropriate instructional strategy.
References


American Educational Research Association, Denver, CO.


Study 2. Developing Shared Epistemic Criteria

This chapter will be submitted to Journal of Research in Science Teaching.
Abstract

Epistemic criteria are the standards people use to evaluate scientific products such as models, evidence, and arguments. In this study, we examined an instructional sequence in which students (1) generated individual lists of epistemic criteria for evaluating scientific models, (2) engaged in discussions in which they collaboratively vetted their criteria, and (3) adopted class lists of criteria. Our analysis indicated that class adopted lists were more sophisticated than the average individual’s list, and four core criteria associated with learners’ understanding of content, epistemic development, and inquiry knowledge were present in the class lists. Learners’ initial shared understanding of the criteria (before discussion) was an important factor for predicting the content of the class lists. We argue that class lists of criteria are sophisticated enough to support student-centered instruction. Implications for scaffolding student-centered, inquiry instruction and developing knowledge-building communities are discussed.
Introduction

Learning environments are often envisioned as communities in which members engage in authentic, disciplinary practices aimed to promote cognitive, social, and epistemic growth (Bielaczyc & Collins, 2013; Brown & Campione, 1996; Scardamalia & Bereiter, 2006). One important aspect of developing learning communities is establishing shared epistemic criteria for evaluating intellectual products. In science classrooms, intellectual products include models, theories, evidence, arguments, and questions (Duschl, 2008; Erickson & Lehrer, 1998; Longino, 2002; NGSS Lead States, 2013; Pluta et al., 2011). Examples of the epistemic criteria scientists use for evaluating the quality of scientific models include (Kuhn, 1977; Longino, 1990; Newton-Smith, 1981; Thagard, 1978):

- Good models have high levels of conceptual coherence and clarity.
- Good models are appropriately parsimonious.
- Good models are consistent with empirical evidence.
- Good models make novel predictions.
- Good models are applicable to current human needs.

Educational research indicates that inquiry instruction emphasizing criteria can promote deep content learning as well as an informed understanding of the nature of science (e.g. Penner, Giles, Lehrer, & Schauble, 1997; Schauble et al., 1995; Schwarz & White, 2005; Smith et al. 2001; White & Frederiksen, 1998). Criteria centered instruction also promotes learning by supporting metacognitive awareness of the goals of instruction, organization and management of learning, and knowledge of how to improve (White & Frederiksen, 2005; Wiliam & Black, 1998). Finally, an instructional focus on epistemic criteria can reliably support the adoption of evaluative norms in the classroom. The Next Generation Science Standards (NGSS Lead States, 2013) emphasize the importance of
establishing evaluative norms around modeling, theorizing, explaining, and engaging in argumentation and working with evidence.

In this article, we investigate learners’ ability to collaboratively propose, vet, and adopt sophisticated class lists of epistemic criteria for model quality. We evaluate the extent to which class lists are suitable for scaffolding science learning. Epistemic criteria have been emphasized in a range of ways in studies examining science learning environments. In some studies, epistemic criteria were never made explicit or formalized during instruction, although an implicit understanding of criteria emerged during inquiry (e.g. Penner et al., 1997; Smith et al., 2001). In other studies, learners were introduced to epistemic criteria through direct, teacher-centered means (e.g. Schwarz & White, 2005).

While evidence indicates that both of these approaches can facilitate learning, we argue that there are advantages to scaffolding inquiry instruction by emphasizing learner-generated, peer-recognized criteria. First, learner-generated criteria may increase engagement and motivation by giving students more choice and ownership over ideas (Blumenfeld, Kempler, & Krajcik, 2006; Rogat, Linnenbrink-Garcia, & DiDonato, 2013). Second, rather than seeing criteria as an impersonal tool dictated by the teacher, student-generated criteria can serve as a personally meaningful tool useful for solving problems. Learners may be more likely to believe that student-generated criteria are valid and worthy of use (Chinn & Samarpungavan, 2001), in part because they may generate reasons for why the criteria are valid. Third, the act of proposing and vetting criteria can serve as an important technique for making students’ thinking visible both to peers and teachers (Collins, 2006). Fourth, the process of sharing criteria may support critical discussion by contributing to increased common ground and shared epistemic framing.
within a community of learners (Scherr & Hammer, 2009; Nathan, et al., 2007; Orsmond, Merry, & Reilling, 1996; Roschelle, 1992). Fifth, the process of proposing, vetting, and adopting criteria is an important scientific practice in its own right; learners who engage in these authentic activities may develop a better understanding of the nature of science (Chinn & Malhotra, 2002; Duschl, 2008). Finally, a focus on student-generated criteria may serve as a reliable method for establishing evaluative norms.

These outcomes are possible if students can collectively develop a reasonable set of epistemic criteria—a set which has similarities to the criteria used by scientists (Kuhn, 1977; Longino, 1994; Newton-Smith, 1981; Samarpungavan, 1992; Thagard, 1978), as well as with criteria found to promote content learning (Schwarz & White, 2005; Smith et al., 2000). If students cannot develop a reasonably sophisticated list of criteria, teacher-centered strategies of introducing criteria may be more productive.

Given the emphasis in recent work on developing community norms (NGSS Lead States, 2013; Ryu & Sandoval, 2012), there is a need for a detailed understanding of the classroom processes by which these norms are generated and sustained. Without a detailed understanding of these processes, teachers will have theoretical knowledge about the value of norms without the practical knowledge of how to use discourse and other tools, such as an emphasis on criteria, to produce them. In this study, we investigated the classroom discussions that are critical to the initial development of norms.

In this article, we examine students’ ability to engage in the process of proposing, vetting, and adopting epistemic criteria for good scientific models with the ultimate aim of promoting evaluative norms. We address three clusters of research questions:

- To what extent can students collaboratively generate a list of class of criteria that can support class learning?
• What discourse processes support and impinge upon the development of a sophisticated list of class criteria? In what ways can instructors support productive discussion about epistemic criteria?

• To what extent do students’ discussions demonstrate adherence to the evaluative norms of epistemic communities?

Before describing the study, we briefly discuss our motivation for focusing on epistemic criteria for good scientific models. We then consider evidence indicating that instruction centered on criteria may promote significant learning gains. Finally, we discuss the role that criteria play in the adoption and adherence to evaluative norms within learning communities.

**Epistemic Criteria for Model Goodness**

Epistemic criteria are used to evaluate the verisimilitude of epistemic products such as scientific models, explanations, arguments, and evidence. Understanding criteria and criteria-related practices is an important part of learning how to participate in science and developing an understanding of the nature of science. Students should be able to differentiate between epistemic criteria and other types of criteria, flexibly apply criteria, understand that scientists sometimes differ in their favored criteria, and recognize that criteria can change or be abandoned in light of criticism by the scientific community (Duschl, 2008; Chinn et al., 2011).

In this study, we focus on epistemic criteria for evaluating good scientific models due to research in the history and philosophy of science (e.g., Giere, 2004; Godfrey-Smith, 2006; Longino, 2002; Nersessian, 2002) and in science education (e.g., Chinn et al., 2008; Raghavan & Glaser, 1995; Schuable & Lehrer, 2006; Schwarz et al., 2009; Schwarz & White, 2005; Smith et al., 2001; Windschitl, Thompson, & Braaten, 2008). Examples of criteria for scientific models include: good models have high levels of
conceptual coherence and clarity, and good models are appropriately parsimonious. As a result of this research, recent recommendations for reforming instruction have advocated placing modeling activities at the center of the science curriculum (National Research Council, 2007; NGSS Lead States, 2013). Drawing on the philosophy of science, we view models as idealizations that scientists use to represent and explain aspects of the world for specific purposes (Giere, 2004, p. 742). Examples of well-known models in science include theories of light such as the wave, particle, and wave/particle models and evolutionary mechanisms such as transmutation, natural selection, and genetic drift. Scientific practice centers on the construction, revision, evaluation and comparison of explanatory models. Epistemic criteria play a significant role in guiding these activities (Kuhn, 1977; Newton-Smith, 1981; Thagard, 1978).

In developing an informed understanding of the nature of science, it is critical that learners learn to differentiate epistemic criteria from the criteria commonly used in traditional school settings. Traditional school criteria focus on students’ effort, aesthetics, or accuracy conceptualized as the match between learners’ work and teachers’, textbook, or testmakers’ conceptions of correctness. These criteria often do not emphasize the explanatory and predictive functions of models and the role of evidence in the construction, evaluation, and critical comparison of models (Chinn & Malhotra, 2002). Some research has indicated that students’ conceptions of models and criteria are more aligned with these traditional school criteria rather than the criteria actually used by scientists (e.g. Carey & Smith, 1993; Grosslight et al., 1991; Treagust et al., 2002). These results are consistent with a much larger body of research suggesting that students’
understanding of the nature of science is nonnormative and unsophisticated (e.g. Lederman, 2007).

There are, however, a few researchers who have come to more optimistic conclusions about students’ understanding of modeling and epistemic criteria. Pluta and colleagues (2011) found that when middle school students’ were provided with a 40- to 50-minute orientation to scientific models, a significant number of normative ideas appeared in student-generated lists of epistemic criteria. Examples include criteria related to the accuracy, explanatory scope, and parsimony of models (cf. Kuhn, 1977).

Similarly, Samarapungavan (1992) found that young learners were able to apply epistemic criteria during a model choice task; however learners struggled to provide explanations for their criteria application. Schwarz and colleagues (2009) identified examples of students drawing on normative epistemic criteria during extended inquiry instruction and Smith and colleagues (2001) found that students who engaged in extended inquiry were able to describe a number of normative criteria for models during interviews conducted after instruction culminated.

In sum, when provided with appropriate support and context, students can demonstrate a reasonably adept understanding of epistemic criteria for model goodness. The present study examines an instructional activity aimed facilitating the development of an appropriate context for using epistemic criteria. Our aim is to develop the instructional support for moving students understanding of criteria forward. Pluta and colleagues (2011) described three categories of criteria: (a) primary criteria, which were defined as being central to the epistemic aims and practices of science (as identified by philosophers of science) and center on the accuracy of the model (the extent to which the
model is congruent with the world in desired respects, as indicated by congruence with empirical observations); (b) secondary criteria, which do not directly impact the accuracy of the model, but contribute to epistemic aims of science; and (c) criteria that are vague or suggest misconceptions about the practices of science. Distinguishing between primary and secondary criteria provides a useful guide for evaluating students’ epistemic cognition during the use of evaluative criteria. It provides a rough indicator of the extent to which students are focused on accuracy of the model versus other model criteria that are important, but can be met by inaccurate as well as accurate models. These three categories are central to the analysis in the present study. We examine instructional strategies aimed at moving students’ focus on secondary criteria (Pluta et al., 2011) to a focus on both primary and secondary criteria (Samarapungavan, 1992; Schwarz et al., 2009; Smith et al., 2001).

It is equally important that students develop an understanding of how criteria are actually used by the scientific community. There is minimal research examining students’ understanding and ability to engage in important aspects of criteria use during inquiry. Students must learn that rival scientists sometimes favor different epistemic criteria and that criteria adopted by the scientific community can be modified, abandoned, replaced, or deemphasized (Kuhn, 1977; Longino, 2002; Toulmin, 1958). For example, early twentieth century American and British geologists preferred focused, localized models, while European geologists (such as Wegener) valued models with broad explanatory scope (Solomon, 2001). Similarly, the history of biology has sometimes seen divergence in the criteria of scientists who give precedence to experimental results and others who give greater significance to careful observations of nature (Mayr, 1982). These
disagreements emerge from a difference in criteria for what counts as the best models and evidence. Having learners’ vet criteria may help them recognize these features of the nature of science. Further, if students can successfully generate and vet criteria on their own, it may also be instructionally appropriate for them to revisit epistemic criteria throughout the school year in order to abandon inadequate criteria and refine and adopt more sophisticated criteria as their understanding of scientific practices matures.

Both applying and understanding how epistemic criteria are developed and shared is a crucial part of understanding scientific practice. While there is growing research on students’ understanding and ability to apply criteria (Smith et al., 2001; Schwarz & White, 2005; Schuable, Glaser, Duschl, Schulve, & John, 1995), there is significantly less research on students’ understanding of the broader issues of how criteria are used within epistemic communities. Finally, there is significant need to develop instructional strategies that reliably support the adoption of evaluative norms, which require the use of epistemic criteria, within the science classroom. We aim to address this gap. An integrated understanding of how to evaluate epistemic products such as models on the basis of shared criteria and how the scientific community evaluates criteria themselves may accelerate the adoption of evaluative norms in classrooms.

Criteria and Learning

There is growing evidence indicating that criteria centered instruction can support content learning, with the majority emphasizing traditional school criteria (cf. Black & Wiliams, 1998; Butler, 1988; Shute, 2008; Zimmerman & Kitsantas, 1997). A smaller corpus of research focuses on epistemic criteria and learning outcomes (Schwarz & White, 2005; White & Frederiksen, 1998). White and Frederiksen (1998) found that low-
achieving learners who were encouraged to reflect on criteria for science inquiry performed significantly better than learners who were not provided with an opportunity to reflect on criteria. In a follow-up study, Schwarz and White (2005) incorporated four epistemic criteria related to the accuracy, plausible mechanism, utility, and consistency of models into an instructional unit; students evaluated a range of models utilizing these criteria. They found that students who used model-criteria demonstrated a better understanding of the nature of modeling and the physics content than students who completed an instructional unit that was identical except it did not emphasize explicit use of criteria (comparison data were drawn from White & Frederiksen, 1998). Results indicated that instruction focused on epistemic criteria can scaffold science learning. However, this study did not provide information on whether students have the ability to generate and use their own criteria, or what students’ own ideas about epistemic criteria are. In addition, Schwarz and White found that most students could not explicitly describe model criteria after instruction, suggesting a weakness in students’ ability to apply these criteria in future settings.

In the present study we aim to evaluate the extent to which learners can collaboratively develop a list of criteria similar to the list provided to students in the study by Schwarz and White (2005) study. Research indicates that students ranging from first graders to university students can collaboratively generate and then apply traditional school criteria that are very similar to their teachers’ standards (Higgins et al., 1994; Orsmond et al., 1996). However, there is little research on students’ ability to collaboratively generate epistemic criteria before instruction. Having students generate their own criteria may support deeper understanding of the criteria they are asked to use.
during instruction and more expansive framing of the learning context—leading to application across environments (Chi & VanLehn, 2012; Engle, Lam Meyer, & Nix, 2012).

**Developing Shared Community Criteria**

Many learning scientists envision classrooms as communities in which learners engage in authentic scientific practices (Brown & Campione, 1996; Scardamalia & Bereiter, 2006). The establishment of communities can be facilitated if students understand and adopt expert epistemic norms. The philosopher Helen Longino (2002) identified four social norms governing model epistemic communities: epistemic communities have (a) public venues for criticism, (b) publicly recognized standards for evaluating intellectual products, (c) members who accept dissent and respond to criticism, and (d) intellectual equality (cf. Goldman, 1999; Kitcher, 1990; Salomon, 2001). These specific norms, all focused on evaluation, have received increasing attention and are aligned with many of the features of model learning communities advanced by education researchers (cf. Brown & Campione, 1996; Pluta et al., 2011; Duschl, 2008; Kelly, 2007; NGSS Lead States, 2013; Scardamalia & Bereiter, 2006). The establishment of learning communities can be facilitated if instructors have reliable strategies to promote student understanding and adoption of these norms. We briefly consider each norm in turn.

A significant focus of research examining the creation and support of learning communities centers on providing students with a venue and resources to engage in constructive evaluation and criticism of intellectual products and ideas (Chinn et al., 2013; Clark et al., 2011; Engle & Conant, 2002; Ryu & Sandoval, 2012). For example,
Zhang and colleagues (2009) documented that knowledge-building discussions were more productive once teachers delegated cognitive responsibility to students. Anderson and colleagues (2001) found that classrooms with open-participation discussion norms were more productive than teacher-guided discussion in supporting the adoption of sophisticated argumentation strategies. Roth and Bowen (1993) described the affordances of an open-inquiry science environment in promoting meaningful learning—in stark contrast to the activity in more traditional teacher-centered school activities. Rosebery, Warren, and Conant (1992) concluded that language minority students demonstrate significant gains in learning and participation when given a venue for constructive peer discussion. While there is significant evidence that developing a venue for critical discussion can lead to powerful learning gains, building and supporting successful communities of learners is challenging.

A number of researchers have documented these challenges (Christodoulou & Osborne, in press). For example, Hogan, Nastasi, & Pressley (1999) reported that teacher-guided discussions were more efficient than peer-guided discussions at attaining higher levels of reasoning and explanation, but that peer discussion was more exploratory and generative. Barron (2003) found that a core difference between successful and unsuccessful inquiry groups was in the extent to which students thoughtfully considered their peers’ ideas. Smith and colleagues (2000) found that compared to students who had engaged in more traditional instruction, students who had received substantial inquiry instruction were able to generate significantly more epistemic criteria during an interview task even though neither the teacher nor students explicitly promulgated the criteria during instruction. However, despite outperforming traditional students, a relatively small
proportion of inquiry students proposed criteria (e.g. 33% of students noted the role of evidence in science). Thus, it was not entirely clear that the entire class widely and publicly adopted the criteria. A critical question we aim to address is how to move students’ reasoning forward within learning communities. We examine whether providing students with an opportunity to collaboratively generate criteria can lead to an instructional scaffold (and criteria themselves) for constructive criticism of intellectual products and ideas. Considering both the optimistic (e.g. Anderson et al., 2001; Engle & Conant, 2002; Rosebery et al., 1992; Zhang et al., 2009) and cautionary (e.g. Hogan et al., 1999; Barron, 2003; Smith et al., 2000) findings, it is difficult to predict whether simply providing students with a venue to propose class criteria will reliably promote the adoption of evaluative norms.

Another critical norm of epistemic communities is the equality of intellectual authority. According to Longino, every member of the community should be regarded as capable of contributing and “the social position or economic power of an individual or group in a community ought not determine who or what perspectives are taken seriously in that community” (Longino 2002, 131). There are a number of ways in which this norm can be violated within learning communities. For example, a subset of students may dominate the development of the community criteria. This can lead to inefficiencies in the learning community—novel individual ideas may not be exposed to the class if some students do not feel empowered to participate. It is also possible that teachers may struggle to give students the authority to generate their own criteria during discussions. Thus, the criteria that result from class discussions may actually represent the teachers’ ideas about criteria. This may undermine the goal of giving students’ ownership of ideas.
This would also be at odds with the scientific community norm of equality of intellectual authority (Longino, 2002; cf. Duschl, 2008).

Finally, within knowledge-building communities there is an acceptance of critical discourse and dissent. In particular, community members’ beliefs and theories must change in response to reasonable criticism. For this to occur within learning communities, students must be able to articulate appropriate explanations, criticisms, and responses to criticisms. There are a number of impediments to quality critical discussion. For example, participants may fall prey to the groupthink phenomenon (cf. Janis, 1972; Levine & Moreland, 1998); groups sometimes focus on quickly reaching consensus rather than proposing and vetting alternate ideas (Goldman, 1999; Longino, 2002). Groupthink and other unfavorable group phenomena could potentially undermine both the development of the best-possible criteria list for supporting learning and impinge upon the development of understanding of core scientific norms. Related to this, we do not have data on the extent to which students will provide justification for their ideas about good criteria. While previous research indicates that students collectively have a rich set of ideas about epistemic criteria (e.g. Pluta et al., 2011) and the ability to apply them (e.g. Samarapungavan, 1992), there are few studies that have succeeded in eliciting students’ justifications for epistemic criteria. Discussing justifications for criteria may lead to deeper appreciation of the rationale and importance of criteria—an important learning goal in science education.

Overall, existing research focuses on the results of extended (e.g. multi-week) interventions and thus it is not clear whether the adoption of evaluative norms occurs gradually over extended periods of time, or if specific instructional activities can reliably
and efficiently promote the adoption of norms. We lack knowledge of both the specific designs that can foster the establishment of norms and of the structure of class discussions that are critical to establishing the norms. Our hypothesis is that if students can collaboratively generate a reasonable set of epistemic criteria this list will serve as a catalyst for developing a community of learners with a robust set of evaluative norms, norms that can move student reasoning forward.

Overview and Goals of the Study

The present study is an extension of our own research examining students’ individual criteria for good models (Pluta et al., 2011). We extend this work by evaluating learners’ ability to collaboratively propose, vet, and adopt epistemic criteria for good scientific models. While there is growing research indicating that criteria centered instruction promotes a range of positive learning outcomes, the field lacks critical information on the types of activities that can reliably contribute to the establishment of norms for evaluating scientific products.

Method

Participants

Participants were 98 learners from four seventh grade science classes taught by two teachers (two classes per teacher) in the same school. The school performed well on New Jersey state math and language arts exams. Both teachers were enthusiastic adopters of model-based, inquiry instruction. In the proceeding school year, both teachers had successfully implemented a yearlong, researcher-developed curriculum focused on model-based science, and thus were familiar with the goals of the focus activity. The 98 students who participated in the present study represent a subset of the 324 students who participated in our previous research (Pluta et al., 2011). We report on four classes to
order to develop a better sense of how replicable aspects of the learning discourse and environment may be across different classes and teachers.

**Instructional Procedures**

First, before the data collection began, students completed a brief 40-50 minute long introduction to model activity. The activity aimed to (a) provide a brief, working definition of models; (b) activate students’ existing knowledge of models, developed through viewing models in textbooks and other media; and (c) present contrasting cases that would help students begin reflecting on specific features of models related to quality. The contrasting cases presented pairs of models on a range of scientific topics that the students were likely to be familiar (e.g. volcanic eruptions, amphibian lifecycles, butterfly lifecycles, climate change). The dimensions along which the models pairs differed included, (a) whether the models were descriptive or explanatory, (b) whether the models were explanatory or representations of data, (c) the degree of model complexity, (d) the presence and type of communicative features (e.g., labels, title), (e) the extent to which the models included details, including both relevant and extraneous details, (f) whether mechanisms were present and, if so, the types of mechanisms, and (g) the extent to which models were consistent with everyday data that students were likely to know about. Additional details of the activity are described in Pluta et al. (2011).

Overall, learners had not received significant instruction or experience with scientific modeling. The data reported reflects learners’ early, preinstructional understanding of criteria and ability to vet criteria.

Second, learners generated individual lists of criteria for evaluating models after viewing and discussing a number of examples of scientific models.
Third, in each class, learners participated in whole class discussions aimed at proposing, vetting, and adopting criteria. The discussions were facilitated by the teachers and resulted in class lists of criteria. Before beginning the activity, learners were told that their class list would be used throughout the course of the year to evaluate, design, and revise scientific models. Teachers were told by the research team that class lists should reflect students’ current thinking about models.

These activities were embedded within each teacher’s normal instructional schedule and took place within the first month of school, immediately after parental consent and student assent were obtained.

**Coding and Analyses**

Learners’ individual lists of criteria, transcribed class discussions, and adopted class lists of criteria were coded according to the Epistemic Criteria for Good Models scheme described in Pluta et al. (2011). The coding scheme identified a target list of criteria for evaluating scientific models. Following the scheme each criterion was subsequently categorized as: (a) primary criteria, which were defined as being central to the epistemic aims and practices of science; (b) secondary criteria, which do not directly impact the accuracy of the model, but contribute to epistemic aims of science; and (c) criteria that are vague or suggest misconceptions about the practices of science (as described above). This scheme serves as a rough guide to the relative importance of criteria for epistemic development.

In all three data sources (individual lists, class discussion, and class lists), students typically presented the characteristics of good models in short phrases, such as—*a good model is neat, detailed, descriptive*. In such cases, we judged that the words in these
phrases reflected different criteria. For example, we coded the word *neat* as an instance of the criterion—good models are organized, which was defined as including neatness. Criteria related to model organization (as well as other criteria focused on communicative elements of models) were categorized as secondary criteria because while many descriptions of scientific practice emphasize the importance of dissemination (e.g., Goldman, 1999) they are less central to the epistemic aims of science. The word *detailed* was classified as an instance of the category—good models are detailed. Ideas about appropriate details and complexity appear closely related to the normative criterion of parsimony (Kuhn, 1977; Popper, 1959; Thagard, 1978), and thus were categorized as primary criteria. The word *descriptive* was coded as an instance of the criterion—good models are descriptive. Although descriptive models do not address the central explanatory goals of science, they have played a significant role in the history of science (e.g., anatomy, botany; Mayr, 1982) and thus this criterion was categorized as a secondary criterion. Additional examples of primary, secondary, vague criteria are described throughout the results section.

In this study, we focus our qualitative analysis on five criteria identified for their instructional and epistemic value: accuracy, plausible mechanism, utility, consistency, and appropriate details. As noted in the introduction, these criteria were emphasized in studies by Schwarz and White (2005) and Schwarz and colleagues (2009) that indicate that students who engaged with these criteria demonstrated significant learning gains. We acknowledge that there are additional epistemic criteria that middle school students should gain familiarity with.
All training and coding scheme development took place on the individual lists of criteria. We reached 85% interrater agreement on 50% of the data for individual lists, 88% interrater agreement on 100% of data for class discussion transcriptions and 91% agreement on 100% of data for class lists. Disagreements were due to clerical errors, oversights, and differences of interpretation; and were resolved through discussion.

In examining the class discussions, we identified conversational turns in which learners attempted to explain criteria or provide reasons for why or why not the class should adopt the criteria. We also differentiated between teacher moves in which teachers did and did not make substantial contributions to the discussion. Interrater reliability was quickly reached at 90% independent agreement on 100% of data. Again, disagreements were resolved through discussion.

**Results**

We first describe learners’ individual lists of criteria, class discussions addressing criteria, and the criteria appearing on class posters. The purposes of this analysis are to orient readers to the overall quality of students’ criteria at each step of the activity and evaluate to the extent to which the discussion promoted adoption of more sophisticated criteria. We then examine the extent to which discussions were aligned with three epistemic norms: (1) intellectual equality, (2) public venues for discussions, and (3) dissent and responses to criticism (cf. Longino, 2002; The fourth epistemic norm, publicly recognized standards for evaluating intellectual products, is aligned with the overall task of collaboratively generating criteria). Finally, we conclude with a qualitative analysis of the class discussions related to the five epistemic criteria that have been identified as important to facilitating the development of students understanding of modeling, inquiry, and content learning (cf. Schwarz & White, 2005; Schwarz et al.,
The purpose of this final section is to examine the extent to which learners’ reveal their understanding of the criteria, and the extent their conceptions match those of scientists.

**Overview of Individual List, Discussion, and Class List Content**

**Individuals’ Criteria.** In aggregate, across all 98 participating students, we identified 31 distinct epistemic criteria in students’ individual lists. We categorized eleven of these criteria as primary; examples of primary criteria included evidential criteria (e.g. “models should be supported by evidence”), explanatory criteria (e.g. “models should say how and why”), and criteria explicitly addressing the accuracy of models (e.g. “the model should be accurate”). We categorized 16 criteria as secondary; these criteria focused on the communicative properties of models and included items such as good models are organized, and constructed with pictures, diagrams, and labels. Finally, we identified four criteria as vague/nonnormative. For example, criteria that stated that models are made of information were categorized as vague/nonnormative criteria, because the term information could suggest that learners believed models were providing static or isolated pieces of information (a view inconsistent to an informed view of science). Pluta et al. (2011) provided a more detailed discussion of the criteria that fall into each of these three categories.

The average individual student proposed 2.1 primary criteria, 3.5 secondary, and .9 vague/nonnormative criteria, for a total of 6.6 criteria. Of the primary criteria, we found that 23% proposed that models are supported by evidence, 40% of all students proposed criteria related to appropriate detail, and 52% of students stated that models are explanations. These three primary criteria align with criteria found to promote students’
understanding of the nature of science and inquiry, and content learning (Schwarz et al. 2005).

When individual responses were aggregated within each of the four classes, we found an average of 8.8 primary, 13.8 secondary, and 3.3 vague/nonnormative criteria were proposed (total 25.8, range 23-29). We did not identify any significant differences across the four classes.

Based on both the overall number of criteria, as well as the number of primary epistemic criteria generated by learners, the data indicates that epistemic criteria are a potentially powerful starting point for instruction aimed towards improving student reasoning and understanding of the nature of science. The distribution of student responses within classes suggests that instruction centered on peer sharing and vetting of ideas related to criteria could be an appropriate instructional strategy (cf. Pluta et al., 2011)—we explore learners’ engagement in just such an activity in the remainder of the article.

**Discussion Criteria.** The four whole-class conversations that followed the task of generating individual criteria lists averaged 92.8 (range = 76-109) conversational turns. An average of 61% of the learners in each class participated in the conversations (range = 57%-65%).

Did students encounter more criteria during the discussion than if they had worked from their individual lists? If students faithfully approached the activity as a knowledge sharing activity, nearly all criteria would be made visible to the class. However, if students engaged in critical evaluation of criteria, because of time
constraints, the discussion may have focused on a limited number of criteria. Further, social dimensions such as shyness or risk aversion could limit the criteria.

Examining the content of the discussions, all instances of criteria that appear in each discussion also appeared on individual lists. On average, each class proposed 20.8 distinct criteria during the discussions. This included an average of 8.0 primary (range = 7-10), 10.0 secondary (range = 9-12), and 2.75 vague/nonnormative (range = 2-3) criteria. An average of five fewer criteria appeared in the class lists than on the average aggregate total from the individual lists (25.8 versus. 20.8). Fewer primary epistemic criteria were proposed during each of the whole-class discussions than appeared on each class’s respective pooled list of individual epistemic criteria (9.8 vs. 8.0). The average proportion of primary criteria proposed during discussions (.38) was not significantly different from the proportion of primary criteria found on pooled lists of individual criteria (.39). The proportion of primary criteria proposed during discussions was slightly greater than the proportion found on the average individual’s list (.32).

These results show that many of the ideas that appeared on individuals’ lists also appeared on the class lists—thus the activity was successful in promoting knowledge sharing. However, the discussions did not encourage learners to distinguish between primary epistemic criteria from other types of criteria. Learners were equally likely to mention secondary and vague/nonnormative criteria found on their individual lists during the discussion.

**Class Poster Criteria.** The criteria appearing on class posters emerged from class conversations; teachers were responsible for recording students’ proposals. Class posters listed an average of 14 (range = 11-16) criteria, showing a further winnowing of criteria
ideas. This included an average of 4.5 (range = 3-6) primary, 7.75 (range = 6-11) secondary, and 1.75 (range = 1-3) vague/nonnormative criteria. All criteria on class posters also appeared on individual learners’ lists. The average proportion of primary epistemic criteria listed on the class lists (.32) was less than the proportion of primary criteria mentioned during discussions and the proportion of primary criteria on the pooled individual lists (.38 and .39, respectively). Four specific criteria were listed on all four of the class lists: (1) evidential criteria, criteria related to (2) clarity and (3) organization of models, and (4) criteria noting that good models are communicated with pictures. We categorized evidential criteria as primary epistemic criteria. The latter three were categorized as secondary criteria.

Overall, the class lists included an average of 4.5 primary criteria. This represents an improvement over the individual lists of criteria in which only an average of 2.1 primary criteria appeared. Thus, the activity did support the sharing of primary epistemic criteria.

Figure 2.1 depicts the overall change in criteria over time. The average individual proposed 6.6 criteria. When individual lists were aggregated, the four classes included an average of 25.8 criteria; this represents the criteria that would be proposed if students were instructed to generate an inclusive list of all student criteria. During the discussions, an average of 20.8 criteria was proposed as part of the vetting process. Finally, the average class list included 14 criteria. The dashed line in figure 2.2 represents an upper bound for the inclusion of primary epistemic criteria in the class discussions and poster. Over the course of the school year teachers could potentially track student progress by evaluating class lists in regards to this upper bound. The dotted line represents a lower
bound for primary criteria. If the number of primary criteria on class lists drops below this level it indicates that class discussion around criteria may not be productive in promoting shared understanding of epistemic criteria.

**Figure 2.1. Changes in the number of recognized epistemic criteria.**

**Discussing Epistemic Criteria.**

The task of proposing, vetting, and adopting criteria emphasizes authentic, dialectical aspects of science (Chinn & Malhotra, 2002; Chinn et al., 2013). The aim was to engage learners’ epistemic resources, and support students’ in differentiating between primary criteria and other types of criteria. In particular, the activity would encourage class members to propose the best ideas from their individual lists, that is, primary criteria; and then, class members would collaboratively discuss and adopt primary criteria class criteria. We now evaluate a number of social and cognitive factors that may have contributed to the content of the discussions, including the adherence to three epistemic norms: (1) intellectual equality, (2) public venues for discussions, and (3) dissent and
responses to criticism. We conclude with an examination of learners’ proposals related to five criteria identified by previous researchers as supporting reasoning, epistemic development, and content learning (cf. Schwarz & White, 2006; Schwarz et al. 2009).

**Intellectual Equality.** We examined whether a select group of learners or the teachers dominated the intellectual content of the discussions; these actors could inhibit the epistemic and cognitive advantages of the collaborative discussions (Longino, 2002).

In looking at student participation, we found that more than half of the learners participated in each of the four classes (on average 61% of learners participated). On average, the most dominant student in each of the four classes participated in 8% of the total turns (the most dominant student overall contributed 11% of the turns). Learners who participated most were typically providing explanations in response to questions or challenges. Overall, the four class participation patterns were consistent with the norm of intellectual equality.

Our instructions to teachers emphasized that their role was to facilitate the discussions rather than be a dominant contributor (cf. Chinn et al., 1995). To examine the teachers’ contributions to the discussions, we categorized their moves in two different ways: Moves that *facilitated* the conversation (e.g. calling on students; asking for clarification and reasons) and moves in which the teacher made *substantial* contributions to the discussion (through stating their own criteria, explaining criteria, providing reasons for criteria). Overall, teachers’ substantial contributions were at the rate of the most dominant students’ participation (an average of 8% of the total conversation). On average, 39% of the substantial contributions made by teachers referenced primary criteria, 59% made reference to secondary criteria, and 33% made reference to
vague/nonnormative criteria. Thirty-seven percent of substantial contributions noted two or more types of criteria, such as when a teacher discussed *models’ explanatory function* (primary) and that models *answer a question* (vague/nonnomative).

Teachers succeeded in facilitating the discussions with a relatively simple array of discourse strategies. Compared to more-traditional teacher-centered discussions, they made relatively few substantial contributions and their contributions were consistent with the norm of intellectual equality.

**Public Venues for Critical Discussion.** We next consider the extent to which class discussions reflected the norm that there is a public venue for critical discussion. Here, we evaluate the extent to which the overlap of learners’ individual ideas contributed to the discussions. If discussions center on the criteria individually shared by a number of learners, it may indicate that learners were proposing criteria that they viewed as “safe” because they (correctly) believed that their peers had also individually generated them. Although this might be helpful for swiftly reaching consensus, it may not be helpful for moving the discussion and ideas forward. If, on the other hand, learners’ discussions centered on a hodgepodge of ideas, unrelated or inversely related to the extent to which individual ideas overlap, it may indicate that learners are sensitive to the epistemic value of dissent and that unique ideas can significantly contribute to knowledge development (Longino, 2002).

We calculated the correlation between the average proportion of individual learners proposing each criterion in the classes and the average number of conversational turns in which each criterion was discussed. The overall correlation was .79, which was
statistically significant ($t = 6.99$, $df = 30$, $p < .001$). The correlation was .80 for primary
criteria, .80 for secondary criteria, and .90 for vague/nonnormative criteria.

We also examined the relationship between the learners’ individual ideas and the
final class posters. The relationship between the proportions of learners who individually
proposed criteria in each class and whether the criteria were adopted was .72 ($t = 5.71$, $df = 30$, $p < .001$).

These results indicate that learners were more likely to discuss shared ideas. To
some degree, this finding mirrors research by Stasser and colleagues (cf. Stasser & Titus,
2003) that demonstrated that group members are less likely to share unique information
that would improve performance on a decision task. However, in the present study, a
number of unique ideas were proposed during discussions, but students’ simply did not
engage with unshared ideas during discussion after the initial proposal. One explanation
for this pattern of results is that the community had not established an associated norm
related to intellectual risk-taking—learners were reluctant to engage with new ideas.
Another possibility is that learners who didn’t already have the idea on their own lists
lacked the epistemic resources to engage with the new criteria. For example, if one
student proposed a criterion related to evidence, other students could not engage around
this criterion because they simply did have sufficient understanding of evidence or related
ideas to engage even though the idea seeded by a peer.

Another approach to this activity would be to allow learners to engage in more
practice with applying their criteria. Anderson and colleagues (2001) found that learners
were more likely to adopt argumentation stratagems after they listen to students use the
strategies in conversation. Thus, the quality and impact of this conversation may be
improved if students are encouraged to provide specific examples of the criteria and demonstrate application to scientific models.

**Explanations and Dissent.** We observed that proposing criteria did not necessarily lead to entirely new or better-articulated criteria because some unique criteria did not get taken up. To what extent did explanations contribute to adoption of criteria? For our purposes, explanations broadly included elaborations, reasons, and examples related to why criteria should be adopted by the class.

Explanations were associated with an average of 64% of the criteria proposed during the four discussions (range = 59% - 67%). When primary criteria were proposed during discussions, learners generated reasons or explanations in support of these criteria at an average of 66% (range = 55% - 71%) of the time. When learners proposed secondary criteria, reasons or explanations were given an average of 68% (range = 75%-66%) of the time. When vague or nonnormative criteria were proposed, learners generated reasons or explanations an average of 50% (range 33% - 100%) of the time.

There was a positive association between the proportion of individual learners sharing an idea before discussion and the number of explanations for each criterion (.68, t = 5.08, df = 30, p < .001).

The results point to the possibility that the extent to which ideas are shared by individual learners before discussion contributed more to whether an idea was simply proposed than for whether learners provided reasons—learners are more fickle about which criteria they provide reasons for.

**Publicly Recognized Standards.** We next evaluated the extent to which discussions (and by extension, learners’ individual ideas) contributed to the final class list
of criteria. We calculated the correlations between the number of conversational turns in which criteria were mentioned and the number of classes that adopted each criterion.

The association between the number of explanations and proportion of classes in which the criteria were adopted was significant (.51, $t = 3.21$, df = 30, p-value = 0.003). The stratified correlations were .36 for primary criteria, .81 for secondary criteria, and .9 for vague/nonnormative criteria. This provides preliminary evidence that quantity of reasons may have influenced whether criteria were adopted for secondary and vague/nonnormative criteria. Quantity of reasons provided did not seem to be associated with the adoption of primary epistemic criteria.

Criteria that promote learning

Schwarz and White (2005) found that middle school students who were introduced to four epistemic criteria (accuracy, plausible mechanism, utility, and consistency) and scaffolded via prompts to evaluate models with these criteria, demonstrated greater learning of physics content and understanding of modeling than students who had completed similar instruction that did not emphasize criteria. However, Schwarz and White (2005) found that after instruction, few learners could not describe these criteria during an interview. Given the central role of criteria in science, we think it is important that learners demonstrate the ability to explicitly name core epistemic criteria. Research on transfer and motivation indicates that understanding and application of criteria will be facilitated if criteria evolve from students’ own ideas than if criteria are provided by teachers through direct, traditional modes of instruction (cf. Anderson et al., 2001; Blumenfeld et al., 2006; Chi & VanLehn, 2012; Engle et al., 2012). Further, a
focus on student generated epistemic criteria may help frame instructional activities towards knowledge building goals (Scherr & Hammer, 2009).

In this section, we examine the extent to which learners discussed the four criteria that Schwarz and White (2005) found facilitated student learning. We also examine a fifth criterion, concerning the representation of appropriate details in models, which is central to more recent work examining growth in students’ model-based reasoning (Chinn et al., 2008; Schwarz et al., 2009).

Accuracy and Evidential Criteria. Scientific accuracy is the degree to which models match target phenomena. Accuracy reflects evidential fit. In our previous work, we categorized accuracy-related and evidential criteria as primary criteria.

We found that an average of 10% of individuals per class-generated criteria related to accuracy (models are true, correct, accurate) and an average of 26% generated criteria related to evidence. In all four classes, at least one student mentioned accuracy-related or evidential criteria during the discussions. Only one class adopted accuracy related criteria on the class posters, while all four classes adopted evidential criteria. In three of the four classes, evidential criteria were proposed by learners relatively early in the discussions (6th, 12th, and 13th student turns). This may indicate that some learners recognized the prominence of evidence in science and modeling.

In examining the data, we identified two themes related to accuracy and evidential criteria. First, while students did propose these types of criteria, these criteria did not lead to in-depth conversations.

In the class in which the accuracy criterion was adopted on the class poster, one students’ evaluation of the criterion suggested a mixed conception of accuracy. After a
student (S1) initially proposed, “facts are correct,” a second student (S2) provided the following reason in support of the accuracy criterion:

S2: I think it's a good idea because if they're incorrect, um, we're still going to get down the right information except the information that you think is right, that you copied down is wrong, and it could like make…and it could destroy an experiment.

On one hand, the student’s response indicates a traditional school-view of accuracy focused on blithely copying information. On the other hand, the reference to potentially flawed experiments suggests that the student might be developing a more normative view of accuracy—perhaps one indicating that scientific progress and knowledge are neither linear nor infallible. In this class, without further discussion, the teacher recorded the criterion on the poster as “Facts should be true based on current knowledge.”

In another class, a student connected ideas about facts (coded as information, 30% of individuals proposed this criterion), evidence (24% of individuals), and accuracy (7% of individuals). In response, the teacher highlighted each criterion’s importance and noted that the class should return to this concept after resolving the discussion about a different criterion:

S1: I was also going to say that all facts that you use and all evidence should be true.
T: Ok, alright, we need to unpack that a bit. Alright does anyone have anything to say against or to add to "answers the question" and "explains the answer?" And then I'm going to come back to the other.

The teacher returned to the topic of evidential criterion by requesting comments on the criterion seven conversational turns later. The student respondent recognized that evidence supports a model, but then presented a new criterion that was only loosely related to evidence: evidence supports a model, as a model is supported by a question (coded as answers a question, 56% of individuals). The conversation once again moved away from evidence:

…(8 turns after S1 proposed that evidence should be true)

T: ...I'm going to come back to "all evidence should be true." Comments about these things? Julie.

S2: Well it's kind of like one of them, but I said "you have to have a supportive question."

T: What do you mean?

S2: On one of these, let's take this one, if you don't have a question up here, that asks you, if it says, "how does a volcano form?" you're not going to want this one.

Again, the conversation quickly diverged from the evidential criterion until a third student mentioned evidence 38 turns later. This student appeared to draw on a pretest assessment that learners completed for our larger research project where learners were
asked to draw arrows between models and evidence (see Pluta et al., 2008 for an example of this type of representation; 11% of individuals proposed criteria mentioning arrows).

...(38 turns after S1 proposed that evidence should be true)

S3: I said if you wanted you could have arrows pointing to like which step is first so that it's not just a bunch of evidence on a piece of paper?

T: Ok, so "use arrows from evidence to model?"

In this class, we have seen how ideas related to evidence emerge in the discussion multiple times: (a) the criterion is initially proposed by a student, (b) the teacher attempts to initiate further discussion of the criterion, and then (c) a student proposes yet another idea related to this criterion. While many students recognized the importance of evidence (26% mentioned evidence on individual lists), the class struggled to have an in depth discussion about the criterion.

In the third class, the evidential criterion emerged during animated vetting of other criteria. Learners initially proposed that models should be constructed with pictures (70% of individuals proposed criteria related to pictures, 22% of individuals proposed criteria related to visuals), but a second student noted that whilst models could be constructed from pictures, they should stay on topic (focus, 43% of individuals). This led to a third student refocusing the conversation on evidence:

S1: Because, um, things are more important than visual aids like you need everything to stick on topic, that's more important than visual aids, so stick on topic should be number one
T: Oh, oh, oh, S1 just stuck in a new concept on us. S1 is throwing out sticking on topic as something that should be important on the list. Wait, wait, wait, guys, guys, shhh. Guys, guys, we've gotta talk about one concept at a time. So we now have this staying on topic going up there. S2, what do you think of that?

S2: I agree, but it shouldn't be number one because I think that, um, diagrams and pictures and explanations are more important than just sticking on topic. Well it's very important…

T: Do you guys think that let's, let's...(students shouting in background)...woah, woah, guys, guys here's the problem: I want to hear from all of you as quickly as possible but if we have this like eruption of talking I lose time that I want to hear what you're brains are thinking. S3?

S3: Well, if you have, if you have good um like text and evidence and good pictures, that won't even matter if the model that they're involved in doesn't relate to the topic because if the topic is like how like the water cycle and like you have a picture of like a volcano erupting but it's a really detailed picture and it's got evidence supporting how it erupts, that's not gonna matter because it doesn't relate to the topic...

A key aspect of this conversation is that the evidential criterion emerged in the discussion about other criteria. This proposal was facilitated by learners’ evaluation of the relative value of different criteria. The evidential criterion was not further considered in this segment of the conversation. However, later in the conversation, a student
proposed, “It needs strong evidence.” The proposal was captured, verbatim, on the poster with no follow-up discussion.

In the fourth class, an evidential criterion was not mentioned until the 19th student turn and the teacher facilitated the use of evidence-focused language. Crucially, the teacher built on a student-proposed, accuracy-specific criterion, “Has accurate information.” One possible reason for why the teacher facilitated the direct use of the term evidence was that an evidential criterion was not mentioned earlier in the class conversations—the teacher might have recognized the important role that this criterion would play in future lessons.

Given the extant research on student reasoning (e.g. Kuhn, 1991; Zimmerman, 2007), we think that it is remarkable that so many students proposed evidential criteria and that they appeared on each class poster. However, congruent with prior research, this was not a criterion that was discussed at length. Of course, it is unlikely that learners had much experience working with evidence in a meaningful way in science classes and so instruction that follows this type of activity would aim to support the development of a more nuanced understanding of evidence through regular experience working with evidence—however a foundation is in place.

**Plausible Mechanism.** Schwarz and colleagues introduced learners to the criterion that good models have plausible mechanisms (Schwarz & White, 2005) or explanatory features (Schwarz, et al., 2009). Because philosophers have identified the generation of mechanistic explanations as a central goal of science (e.g., Giere, 1988; Kitcher, 1993), we classified this criterion as a primary epistemic criterion. On average,
52% of individuals in each class proposed criteria related the explanatory function of models (with no significant differences between teachers).

However, we found that only two of the classes (same teacher) discussed and adopted explanatory criteria on the class list. The discussions resulted in the following class criteria: “Explain the answer using words, pictures, diagrams,” “Explains how and why,” “and Explains everything well.” Although these criteria can be further explicated as learners gain more experience with models, explanatory criteria were central to the class lists.

In the second teachers’ classes there was only one turn in which “explanation” was explicitly mentioned:

S: I agree, but it shouldn't be number one because I think that, um, diagrams and pictures and explanations are more important than just sticking on topic. Well it's very important…

Neither this student nor his peers fleshed out the explanation component of this proposal. However, a positive development was that the student weighed explanation, as well as two other somewhat model-specific criteria (we think criteria related to diagrams and pictures are model-specific because we do not think that there are many instances in which middle school learners utilize pictures or diagrams to represent phenomena in school, cf. Chinn & Malhotra, 2002) as more important than a general well-established, school criterion (“sticking on topic” was categorized as model focus). In examining the transcripts of this teachers’ class, we also found that some of the models that learners discussed as examples were explanatory (e.g. models of *how* volcanoes erupt, the water cycle). One possibility for the lack of focus on “explanation,” even though it appeared on
many individuals’ lists is that the idea that models are explanatory was implicitly understood or definitional and thus not an obvious target for discussion.

We have argued that ideas about sequence and steps also invoke ideas about mechanism, and explanation, all key elements of scientific models (Pluta et al., 2011; cf. Machamer et al., 2000). An average of 32% of learners per class proposed criteria related to sequence on their individual lists. Criteria related to sequence were discussed in all four classes and adopted in three. However, there were no instances in the discussions in which learners or teachers made explicit connections between sequence and mechanism or explanation. Here is an example of a typical proposal in which sequence-related criteria were proposed:

S1: The model should be realistic…like if…[inaudible]
T: I'm not sure what you mean by that.
S1: If you were doing like how a butterfly, like the frog one if you were doing that you wouldn't say that the frog turns into or butterfly or like something
T: Ok, alright. Julie.
S2: Um, well you have to have a good model you, like when you have something that leads up to it like if for going to school if that were something you were to do you're going to have to do something that leads up to it like how you get to school.

Overall, more than half of students in each class proposed criteria related to plausible mechanism. Two of the classes successfully adopted criteria related to plausible
mechanism. Criteria related to sequence were adopted in three classes. Overall, the classes established a foundation for engaging in model-based reasoning.

Utility. Criteria related to model utility concern the usefulness of models to scientists or science learners (Schwarz & White, 2005). We found that 42% of learners noted that models would only be useful if they “stayed on topic.” Further this idea was mentioned in all four discussions and was adopted on three class criteria lists. In general, this criterion appears to be the type that is typically found in traditional school settings. However, we think that with practice students could flesh out the epistemic purposes of this criterion.

Twenty-percent of individuals noted the importance of thinking about audience when evaluating models on their individual lists. In the discussions, learners noted how different audiences needed models to be specified in different ways, for example:

S: Yeah, so, it, I, it's kind of pointless arguing about this because you can't, it really only matters who you're talking to, like if you're talking to a scientist who might be using really big words…maybe someone who's like not a like not a scientist or something you might talk like a...beginner, 'cause like things…like this tells you the simple steps…so people who aren't...

Criteria related to audience were proposed in three discussions and adopted in two classes (one class for each teacher).

There were at least two important criteria related to the goals of models. As we discussed above some learners and classes believed that models should have an
explanatory function (two classes adopted this type of criteria). Another type of criteria was related to the descriptive goals of scientific models. We found that, on average, 28% of individuals proposed this criterion, it was proposed in three classes, and adopted in two. The same two classes that adopted the explanatory criteria also adopted criteria focused on the descriptive goal of models. Moving forward, we would predict that these learners would have a better understanding of the different types of models.

**Consistency.** An internally consistent model is one that does not contain mutually inconsistent propositions (Samarapungavan, 1992) Although criteria related to the consistency of models was a criterion presented to learners in their instructional intervention, Schwarz and White (2005) did not comment on learners’ understanding of consistency in their article, thus we assume that the learners did not mention this type of criteria in their study’s interview component. In the present analyses we also did not identify any instances related to the consistency of models.

However, in an interview study, Samarapungavan (1992) found that seven- and eleven-year olds preferred logically consistent theories over inconsistent theories (specifically, theories which were inconsistent with the children’s existing beliefs). But, like Schwarz and White, Pluta et al., and the present study, students in Samarapungavan’s study were not successful at articulating this criterion. Overall, it seems that although learners may be able to apply criteria related to internal consistency, it is not particularly salient to them without an explicit context. Given these findings, we think that criteria related to consistency may be a good target for more direct instruction to support learners’ explicit understanding of this criterion.
**Appropriate Details.** Schwarz and colleagues (Schwarz et al., 2009) have focused on examining learners’ understanding of appropriate details. In our previous work (Pluta et al., 2011), we categorized criteria related to “appropriate details” and “appropriate complexity” as primary criteria because they are closely related to parsimony, a normative criterion (Kuhn, 1977; Popper, 1959). In addition, philosophers have discussed how models connect to the world in appropriate respects and to appropriate degrees; models with appropriate detail and complexity meet these purposes (Giere, 2002; Longino, 2001).

For complexity related criteria, an average of 8% of individuals generated this type of criterion, it was proposed in two discussions, and adopted in one class. For detail related criteria, an average of 46% of individuals generated this type of criterion, it was proposed in four discussions, and adopted in three.

We found that these two criteria were a major focus of the discussion within a single class, with these two criteria being directly mentioned in 25% of the conversational turns. One student proposal was that less details simply makes the model more comprehensible:

S: Yeah, because if you put too much detail in it, people are gonna be like, alright I'm just hearing the same thing over and over again in different words and you're just gonna get bored. You need a, you need like, you need to skip to the point, you need to give like just enough detail and then give the conclusion you know. Otherwise it just doesn't make, it just doesn't make sense to add too much detail.
A new student argued that readers (without specifying who the readers are) prefer and will choose more detailed models, but without explaining exactly why this would be the case:

S: Yeah, but a lot of detail gives them a bigger chance of picking your model.

On one hand, it is notable that the student recognized that model choice is a central aspect of scientific practice. However, this could be interpreted as use of established school-perspective by the student; such as, teachers and perhaps peers prefer products that are more detailed when grading (and indeed evidence suggests that graders of standardized tests tend to give lengthier essay’s more credit; in science, we suspect that a “traditional” teacher might prefer a more detailed model because it is evidence of hard work—even if the details are unnecessary or inaccurate). Regardless, it is a claim that deserves to be fleshed out within classroom discussions.

The initial learners’ reply was that some extraneous details simply would not make a difference to understanding. Assuming that the student means understanding of the key phenomena, she seems to have a normative view of appropriate detail (another student also agrees with that claim):

S1: Yeah, but if you put enough detail in it, you'll understand it the same way as putting in over detail.

S2: That's what I'm saying, now you're agreeing with what I'm saying.

A few turns later, a student noted that quality of detail matters:

S3: About the details that they were talking about, it's not necessarily how much detail you add in because you can add in a lot of details but it just
matters how good the details is, you can have like two sentences of amazing details and then like five sentences, like a paragraph of like tons and tons of detail but the two sentences will be better because it will be more descriptions.

S4: But if you understand it, why does it matter if it's so much detail?

Also in this conversation was the point addressed above: the extent to which details should be represented depends on the audience. Specifically, a scientist requires more detail than a middle school student. As the examples from learners’ discussions on appropriate detail indicate, learners can carefully vet criteria and do in such a way that the teacher can step-back from a whole-class conversation.

**Discussion**

The results of this study indicate that learners can collaboratively identify a list of criteria that is more sophisticated than the average individual’s list and contains many of the criteria that appear on the criteria lists generated by science educators. In particular, learners proposed and identified instructionally valuable epistemic criteria related to (a) accuracy, (b) mechanisms, (c) utility, (d) and appropriate details of models. The discussions generally adhered to Longino’s (2002) norms for knowledge-building communities. First, participation patterns did not indicate a violation of the norm of intellectual equality. Neither students nor teachers dominated the conversations. Second, we found that reasons and number of conversational turns centered on criteria were associated with class adoption of criteria, suggesting that students did not violate the norms for critical discussion and dissent. Further, although there is room for significant improvement, students did demonstrate the ability to justify many of their criteria
proposals. That said, discussions did tend to focus on criteria already shared by many individuals, rather than on more unique criteria. Developing a norm centered on intellectual risk-taking would likely benefit class conversations.

The findings provide evidence of one effective design for establishing scientific norms in the classroom—with the norms instantiated in this case as criteria for model quality. The design involves engaging students individually and in pairs to consider which of two pairs of models is better, and then asking each individual student to write their thoughts about the characteristics of good models. At this point, both teachers in this study were successful in leading class discussions that established a set of class criteria for evaluating model quality. The key discourse moves are quite simple: (a) calling on students, (b) asking for clarification, and (c) asking for reasons.

These results are interesting because they provide initial evidence that students, before significant instruction or practice, can collectively generate a list of epistemic criteria that can be used to support the development of evaluative norms. We will now unpack the design implications of these findings and propose future research objectives.

**Design Modification to Introductory Tasks**

Overall, students were collectively successful in including a number of core epistemic criteria on the class lists. How can the instructional design be improved to move student discussions forward?

A number of important criteria appeared on students’ individual lists and in discussions, but did not appear on the final class lists. For example, the criteria related to explanation were not explicitly recorded in two classes. In the current instantiation of the instructional activity we did not provide teachers with guidance on which criteria to
encourage students to include on the class lists. In future designs, we will provide explicit instruction to teachers about which epistemic criteria would be most important for engaging in the modeling activities that will follow this introductory task. Another modification to the current design will be to provide teachers with time to view students’ individual criteria before the discussion. This will give teachers an opportunity to identify important criteria so that they can seed criteria that have not been proposed during discussions. For example, the teacher could say something along the lines of: “I noticed that Bobby made the suggestion that ‘models are supported by evidence’—what do you guys think?” We believe these simple additions to the teacher manual will significantly improve teachers’ facilitation of the discussions.

A related set of research questions include, (a) what criteria will teachers with different levels of training and experience identify as most important to promoting model-based inquiry, and (b) to what extent can teachers extrapolate from core criteria (e.g. mechanism and explanation) to related criteria (e.g. sequence). In the present study, we found that teachers’ utterances were centered on secondary criteria and hypothesized that these teachers with researcher guided experience in model-based instruction may have been aiming to move the conversation toward more substantive criteria.

We found that class discussions focused on the criteria that were most prevalent on individuals’ lists. We hypothesized that students had not adopted an intellectual risk-taking norm. In our redesign we will work to promote this norm. This may take the form of guidance within the introduction to model task or through explicit instruction and discussion facilitated by teachers (presented in our teacher manual and professional development).
We also identified a few important criteria that did not appear in the individual or class lists. For example, students did not generate criteria related to model consistency; consistency was included in Schwarz and colleagues (2005, 2009) lists. There are also a number of criteria identified by philosophers that students did not generate, such as criteria relating to the internal and external consistency of models (Kuhn, 1977), observational nesting, the historical track record of empirical success, and the smoothness with which adjustments can be made in the face of explanatory failure (Newton-Smith, 1981). Based on the results of this study, we will likely redesign the introduction to model activity to include additional model comparisons, such as comparisons around consistency, for students to consider before generating individual and class lists. We will assure alignment between the introduction to models activity, the teacher manual, and model-based inquiry instruction that follows this activity.

**Future Instructional Design**

Overall, the activities described in this study provide a foundation for student engagement in sustained model-based inquiry. On the basis of this work, we have begun aligning instructional modules the results of the class discussions and lists.

First, on the basis of student discussions and class lists, we have incorporated modeling activities that will allow students to apply the criteria on their class lists. This will include explicit prompts for students to provide justifications for their criteria use. Eventually activities will be designed to encourage students to refine their understanding of core criteria. These activities will have a significant focus on collaborate evaluation of models (cf. Chinn et al., 2008; Pluta et al., 2008).
Second, given the success of the class discussion around criteria, we will incorporate regular opportunities for students to reevaluate and modify their class lists of criteria. This will encourage student to make their maturing thinking about models visible to teachers and peers. This will also help students recognize that criteria (like models) can change or be abandoned in light of criticism by the community. A critical question will be how explicit will these activities need to be in order for students to recognize that their patterns of activities reflect those of scientists.

In general, class discussions adhered to Longino’s norms. We will also develop strategies to assure that teachers and students monitor continual adherence and measure whether students’ are developing an explicit understanding of these norms.

Summary

Many reports on the development of inquiry learning environments either are silent on how epistemic criteria are established and what the criteria actually are, or teachers’ provide criteria at the outset of instruction and often are responsible for holding students’ accountable to these criteria. In this article, we provided initial evidence that middle school students can engage in conversation around epistemic criteria, and that the conversation contributes to the sharing of important criteria. We provide important information on which collaborative classroom activities move students’ thinking about epistemic criteria forward.
References


Study 3. Coordinating Evidence: Learners’ Adaptive Strategy Use

This chapter will be submitted to Science Education.
Abstract

We report on the results of an interview study that investigated learners’ ability to coordinate and reconcile the results of conflicting scientific evidence. Twenty-nine middle school students read pairs of studies bearing on four scientific problems and responded to six questions aimed at providing insights into their strategies for coordinating evidence and epistemic cognition. We identified ten unique reasoning strategies learners used to coordinate the results of studies. The four most common strategies were (a) identifying methodological flaws, (b) identifying design differences, (c) subtyping the subjects of the studies, and (d) proposing more inclusive explanations. Learners applied these and other strategies adaptively across four different problems. Overall, learners’ patterns of reasoning were consistent with having sophisticated epistemic beliefs about the nature of science. We argue that learners’ epistemic cognition is more sophisticated than previous research suggests. Implications for improving research and instruction on reasoning and epistemic cognition are discussed.
Introduction

When developing or choosing explanatory models of phenomena, people regularly encounter evidence that is ostensibly in conflict but on closer inspection illuminates more complex relationships such as causal chains, interactions, mechanisms, and subtypes. For example, as scientists have learned more about cholesterol, the public faces contradictory reports on whether it is harmful or beneficial, and whether it is confused with other causes of heart disease. Scientists have discovered that there are multiple subtypes of cholesterol and that there are relationships between the subtypes and personal and contextual variables such as stress and health outcomes. Additional research, which appears to contradict prior research at first glance, may not contradict earlier research at all. Rather, new research may point to additional complexities. It is important for people to become adaptive thinkers, ready to anticipate and accommodate new evidence and learn about complex models that can integrate diverse evidence.

The reasoning skills required to successfully interpret and evaluate scientific evidence and models are supported by the reasoner’s epistemic beliefs and beliefs about the nature of science (Khishfe & Abd-El-Khalick, 2002; King & Kitchener, 2004; Kuhn, Cheney, & Weinstock, 2000; Pluta, Chinn, & Duncan, 2011; Sandoval, 2005; Weinstock & Cronin, 2003; Zeineddin & Abd-El-Khalick, 2010). Consider reasoners’ beliefs about the complexity of knowledge, which may range from seeing knowledge as an accumulation of piecemeal facts (simple knowledge) to viewing knowledge as highly interrelated concepts (complex knowledge): Reasoners who believe that scientific knowledge is complex may employ multiple reasoning strategies (e.g. identifying possible methodological flaws, coordinating details, generating alternative explanations, searching for additional evidence) in order to understand the complexity of a problem.
Reasoners who believe that knowledge is simple may only apply a single strategy, believing that a single strategy suffices. Similarly, reasoners who appreciate the role of evidence, as well as the criteria for evaluating evidence, will likely approach scientific problems in a very different way than reasoners who do not appreciate the importance of evidence in knowledge construction (cf. Hofer, 2004; Muis, 2008). The latter type of reasoner (if it is fair to call them reasoners at all) may blithely ignore or dismiss evidence (Chinn & Brewer, 1993). It is important for researchers to understand how epistemic beliefs and commitments are manifested during reasoning.

We report an interview study that investigates middle school students’ ability to coordinate evidence from pairs of conflicting studies. We describe the patterns of reasoning that emerge across different scientific problems in order to develop a deeper understanding of epistemic cognition. This research extends the extant research on epistemic cognition (EC) and reasoning in four ways. First, existing research is relatively silent on how learners actually reconcile conflicting viewpoints by drawing on evidence. In most studies, participants are not provided with evidence in which details about the methodology can be evaluated. We aim to describe students’ abilities at this important scientific reasoning skill. Second, many studies examining EC predominantly utilize context-free measures, such as questionnaires or interview protocols, in which respondents are provided with limited or no information about the topic; researchers utilizing context-free assessments often argue that students’ EC is unsophisticated (e.g. Allchin, 2012; Hofer & Pintrich, 1997; King & Kitchener, 2004; Kuhn et al., 2000; Lederman, 2007). We aim to explore if providing learners with contextual information during assessment, such as detailed evidence, may elicit critical information about
learners’ epistemic beliefs and commitments. If EC is topic-specific or if providing respondents with contextual information significantly alters their responses, it calls into question the standing interpretations of existing measures of EC. Third, a significant amount of EC research aims to assess epistemic beliefs by having respondents evaluate general statements related to potential beliefs (Muis, Bendixen, & Haerle, 2006). In this article, we argue that patterns of reasoning across problems and topics are closely aligned with epistemic beliefs, and thus may provide important information about learners’ beliefs. Finally, there has been a strong focus on domain or discipline specific EC (e.g. beliefs about science or history); domain-specific measures often have poor psychometric properties (DeBacker, Crowson, Beesley, Thoma, & Hestevold, 2008) and often only a moderate relationship to learning outcomes (Muis et al., 2006). We examine the extent to which it is efficacious to focus on topic-specific aspects of EC (e.g., within science, beliefs about evolutionary mechanisms, cell organelles, dinosaur metabolism). A few recent studies suggest that topic-specific epistemic beliefs may better predict student performance than domain-general or domain-specific beliefs (Kienhues, Bromme, & Stahl, 2008; Porsch & Bromme, 2011). We aim to provide an explanation for why this may be the case.

Overall, this study addresses four predictions based on the extant research:

Prediction 1. Students will have few, if any, strategies for coordinating evidence.

Prediction 2. Students’ patterns and targets of reasoning will not be consistent with sophisticated epistemic beliefs.
Prediction 3. Providing students with contextual information, such as evidence, will have minimal impact on their reasoning and epistemic cognition.

Prediction 4. Students will demonstrate domain-level epistemic beliefs.

Before describing the study, we survey the research underlying these predictions. In the introduction, we first flesh out what the existing research suggests about sophistication and specificity of learners’ EC; we then consider how epistemic beliefs are manifested in reasoning. The body of this study focuses on describing students’ reasoning about conflicting scientific evidence. We conclude by using our description of students’ reasoning to examine the four predictions and draw inferences about the sophistication of students’ epistemic cognition.

Epistemic Cognition

Epistemic cognition refers to thinking about epistemic or epistemological matters (Chinn, Buckland, & Samarapungavan, 2011; Greene, Torney-Purta, & Azevedo, 2010; King & Kitchener, 2004). A significant amount of scholarship has focused on learners’ epistemic beliefs about the complexity, certainty, source, and justification of knowledge (cf. Hofer & Pintrich, 1997; Muis et al., 2006). Under the guise of Nature of Science (NOS), science education researchers have identified and examined additional, science-specific aspects of EC, such as knowledge (or beliefs) about, (a) distinctions between theories and laws (b) theory-ladenness of science, (c) specific empirical features of science, and (d) creative and imaginative aspects of science (cf. Lederman, 2007). Over the past twenty years, EC has received sustained attention because of its relationship to
learning processes and outcomes. For example, Qian and Alvermann (1995) found that learners who believed that knowledge is complex and uncertain were more likely to restructure their beliefs on a conceptual change task. Songer and Linn (1991) found that a more informed understanding of the nature of science was associated with science learning. Greene, Muis, and Pieschl (2010) have examined the relationship between EC and self-regulated learning and Bråten, Britt, Strømsø, and Rouet (2011) summarized the connections between learners’ epistemic beliefs and the strategies they used as they learned from multiple documents. In addition, sophisticated beliefs may foster an appreciation of epistemic practices that encourages and motivates students to (enthusiastically) participate in epistemic communities, such as the science community (Lederman, 2007).

In this article, EC encompasses the way learners’ patterns of reasoning and explanations reflect their epistemic beliefs. Because this study focuses on science, epistemic beliefs should be viewed as being synonymous with beliefs about the nature of science.

**Sophistication of Epistemic Beliefs**

A number of researchers have constructed models that describe the development of epistemic beliefs. One approach has been to characterize epistemic development as a progression through a series of stages (e.g. Carey & Smith, 1993; King & Kitchener, 2004; Kuhn et al., 2000). Developmental approaches assume that learners’ epistemic beliefs are relatively coherent and undergo qualitative shifts. Epistemic growth is characterized as a progression from (1) realist, seeing knowledge as a direct copy of reality; (2) absolutist, believing knowledge is absolute and processes of knowing are
objective; (3) multiplists or relativists, believing knowledge is subjective and relative and, therefore, indeterminate because of multiple points of view; to (4) evaluativist, where the acceptance of subjective and objective aspects of knowledge permits a degree of evaluation and judgment of knowledge claims (Kuhn et al., 2000; cf. Carey & Smith, 1993; King & Kitchener, 2004, for comparable models). Research suggests that few learners develop evaluativist-level cognition. Kuhn and colleagues (2000) found that only 16 to 40% of eighth graders and 25 to 45% of undergraduates could be categorized as evaluativists (varying by domain). Kitchener et al. (1993) reported that only between 25 and 50% of undergraduate aged students (ages 18 to 22) and 0 to 5% of middle school (and younger high school students) demonstrated evaluativist thinking (King and Kitchener called this stage reflective thinking). Focusing on students’ understanding of NOS, Carey, Smith, and colleagues (Carey & Smith, 1993; Grosslight, Unger, Jay, & Smith, 1991; Smith, Maclin, Houghton, & Hennessey, 2000) concluded that middle school science students did not demonstrate a sophisticated understanding of scientific knowledge before instruction (they dubbed this Stage Three thinking, characterized by a recognition of the “cyclic, cumulative nature of science and identifies the goal as the construction of even deeper explanations of the natural world,” Carey & Smith, 1993, p. 251). In sum, most developmental approaches to studying EC have found that students’ EC is impoverished.

Other researchers have focused on distinct dimensions of epistemic development—these dimensions are assumed to develop somewhat independently. Much of this work has concentrated on students’ beliefs about knowledge, which encompasses beliefs about (1) certainty and (2) complexity, and beliefs about knowing, which
encompasses beliefs about the (3) source and (4) justification of knowledge. Most research suggests that learners’ (college and grade school students) beliefs about knowledge and knowing are inconsistent with normative perspectives (Hofer, 2001; Khishfe & Abd-El-Khalick, 2002; King & Kitchener, 2004; Kuhn et al., 2000). For example, in examining instructional practices aimed to improve individual dimensions of NOS knowledge, Khishfe and Abd-El-Khalick (2002) concluded that before instruction only 6% of sixth graders had an informed view about the tentativeness of scientific knowledge (which mirrors the certainty dimension), 9% could distinguish between observation and inference (which mirrors the justification of knowledge dimension), 6% had an informed view about the empirical basis of science (justification of knowledge), and 3% articulated informed views of the creative and imaginative NOS (mirrors the certainty dimension). After explicit instruction on NOS knowledge, students’ articulated understanding of each of these dimensions increased to 52%, 40% 48%, and 34%, respectively. Based on this research there is little reason to believe that middle school students, the focus participants in the present study, should be able to demonstrate sophisticated epistemic beliefs when evaluating evidence.

Overall, there is significant evidence suggesting that students’ (and laypeople’s) epistemic cognition is unsophisticated without extensive instruction. In fact, some studies suggest that even teachers’ epistemic beliefs are deficient (Olafson & Schraw, 2006; Peters-Burton & Baynard, 2012; Windschitl & Thompson, 2006). The focus of the present study is on middle school students’ EC. A prediction based on the existing evidence is unequivocal: few middle school students will exhibit sophisticated epistemic beliefs.
Specificity of Epistemic Beliefs

The extent to which students’ epistemic beliefs are nuanced and specific to domains has received significant attention by researchers. Understanding the specificity of EC has important implication for addressing the relationship between EC, learning processes and outcomes, and informs the design of EC measures (Muis et al., 2006; Schraw, 2001). Initially, EC research focused on students’ global beliefs, which were thought to generalize across academic domains and topics (domain-general beliefs; cf. Schommer, 1990). However, a corpus of research has converged on the conclusion that students hold domain-specific beliefs. In a review, Muis et al. (2006) identified 17 studies that supported the existence of different beliefs sets corresponding with specific academic domains, such as history, psychology, science, and mathematics. For example, Greene and colleagues (2010) identified differences in beliefs between math and history domains.

In examining the role of refutational texts on conceptual change, Mason, Gava, and Boldrin (2008) focused on epistemic beliefs about science. They identified a modest interaction between epistemic beliefs, topic interest, and instructional approach in promoting conceptual change. Science education researchers, of course, focus exclusively on students’ beliefs about the nature of science (Lederman, 2007). For example, Khishfe and Abd-El-Khalick (2002) compared explicit instruction addressing the nature of science to inquiry approaches to implicitly introducing students to the nature of science. Their results suggested that explicit instruction best supported the development of informed beliefs about science. Overall, most EC research focuses on domain-specific beliefs as a predictor of learning processes and performance, or as an outcome of instruction. However, relationships between EC measures and learning processes and
learning outcomes are often only marginally significant, if significant at all (cf. Mason et al., 2008; Muis, 2008; Strømsø, Bråten, & Samuelstuen, 2008)

Recently, there has been a growing interest in examining learners’ topic-specific or problem-specific beliefs. Topic-specific beliefs may be a better predictor of learning processes and outcomes. In science, it would be reasonable for a student to demonstrate absolutist-level thinking about problems related to basic human anatomy (e.g. location and function of organs or structures), but perhaps demonstrate evaluativist thinking for cutting edge physics problems (e.g. speed of neutrino particles). It is extremely unlikely that scientists would decide that the stomach has an entirely different function than it is believed to have now. It is more likely that important elements of physics knowledge would be revised. Even middle school students could be sensitive to these types of differences in certainty evaluations. It is important to understand the extent to which students’ epistemic beliefs about a topic are normatively calibrated. We briefly consider evidence related to the topic or problem specificity of epistemic beliefs in students.

Trautwein and Ludtke (2007) utilized questionnaires to examine the specificity of university students’ epistemic beliefs about the certainty of knowledge, across ten science topics (e.g. cholesterol causes heart attacks, theory of relativity, computer games involving violence lead to higher aggression). Their results indicated that students’ beliefs about certainty varied across topics. In examining historical epistemic beliefs, Gottlieb and Wineburg (2012) compared the ways in which religious historians and other experts examined historical documents related to Biblical Exodus and the origins of the American Thanksgiving holiday. Religious historians engaged different epistemic criteria for each set of documents and were explicitly aware that their two sets of epistemic
beliefs were in conflict. One interpretation of this result is that religious historians had topic-specific epistemic beliefs about history.

Instructional studies also provide insights into the specificity of epistemic beliefs. Kienhues, Bromme, and Stahl (2008) examined the extent to which different forms of instruction were associated with shifts in epistemic beliefs. Students read brief refutational or expository texts about DNA fingerprinting. It was hypothesized that refutational texts would promote epistemic doubt, thus inducing belief change. Kienhues et al.’s (2008) results suggested that either students’ beliefs about the nature of science shifted after the brief intervention or that they engaged a different set of beliefs. They concluded that it may be possible to improve domain-specific (scientific) epistemic beliefs or that epistemic beliefs are simply unstable. Another interpretation is that students had existing topic-specific epistemic beliefs before instruction; once students were provided with topic-specific information, they were more likely to draw on their topic-specific beliefs than on their general beliefs about science. There are a number of other instructional studies reporting increasing sophistication in EC that could be interpreted as showing that students are more likely to draw on topic-specific beliefs rather than general-level beliefs after instruction (e.g. Conley, Pintrich, Vekiri, & Harrison, 2004; Hogan & Maglienti, 2001; Khishfe & Abd-El-Khalick, 2002; Metz, 2011; Smith et al., 2000). What is notable about the Kienhues et al. (2008) study is that such a brief instructional activity encouraged students to demonstrate a different set of beliefs. Overall, we think this new evidence indicates that the improvement in EC described in most instructional studies is due to the instruction providing both a context for responding to EC measures in a topic-specific way and supporting some tangible
development in students’ overall epistemic beliefs (topic-specific, domain-specific, and
global).

Finally, there is emerging evidence that EC is sensitive to contextual features and
information. Hammer and colleagues have drawn attention to this possibility through a
number of case studies (e.g. Louca, Elby, Hammer, & Kagey, 2004; Scherr & Hammer,
2009). Porsch and Bromme (2011) have generated compelling experimental data. They
manipulated secondary school students’ responses to EC questionnaires by highlighting
either the static and consensus aspects or conflicting and controversial aspects of
explanations of coastal tide mechanisms. Learners’ who were introduced to conflicting or
controversial aspects, (a) demonstrated more sophisticated epistemic beliefs on traditional
questionnaire measures, (b) noted that “credible” sources should be consulted, and (c)
said that more sources should be consulted when learning about the topic. We think that
this study suggests that students’ epistemic commitments are influenced by contextual
clues provided by the experimenters. We aim to contribute to this literature related to the
context sensitivity of EC by generating an explanation for why EC is influenced by
contextual information.

Researchers’ understanding of the specificity of EC is evolving. There have been
only a few studies examining topic-specific EC in undergraduates and experts. The
dominant view, supported by the preponderance of evidence, is that domain-specific
epistemic beliefs have the most utility in predicting learning processes and performance
(cf. Greene et al., 2010; Muis et al., 2006). The present study examines middle school
students’ EC and reasoning across four science topics. A prediction based on the extant
research is that the students should not demonstrate topic-specific beliefs on the four
science topics, but rather demonstrate consistent, domain-level beliefs. The present study has the potential to provide more clarity to the fields’ theoretical conceptualization of epistemic cognition.

**Epistemic Cognition and Reasoning**

“At the heart of the evaluativist epistemological position is the view that reasoned argument is worthwhile and the most productive path to knowledge” (Kuhn et al., 2000, p. 35). “Reflective thinkers consistently and comfortably use evidence and reason in support of their judgments (King & Kitchener, 2004, p. 9). We view reasoning as an important type of EC and a critical indicator of epistemic beliefs. In this section, we discuss how the targets and patterns of reasoning reveal epistemic beliefs. Examining both the targets and patterns of reasoning may yield new information about EC, and are critical for examining the development of epistemic beliefs during learning activities (Scherr & Hammer, 2009).

**Reasoning about Evidence**

The majority of research aimed at assessing EC has utilized context-free questionnaires that do not require participants to engage in actual reasoning. Students and instructors are typically asked to reflect on reasoning and evidence-based practices by responding to items such as: (a) Once scientists have the result of an experiment, that becomes the only answer, (b) Scientists always agree about what is true in science (Mason et al., 2008), (c) I teach my students to provide evidence for their thinking (Hennessey, Murphy, & Kulikowich, 2013). Many of these types of measures have been criticized for poorly conceptualizing epistemic dimensions and fail to meet reliability and validity criteria (e.g. Chinn et al., 2011; DeBacker et al., 2008; Greene et al., 2010; Muis
et al., 2006; Murphy, Edwards, Buehl, & Zeruth, 2007). Due to these problems, the results of studies utilizing these context-free measures should be viewed with caution.

Another approach to evaluating EC is to engage students in reasoning activities. King and Kitchener (2004) assessed EC by utilizing an interview task in which participants responded to ill-structured problems. Reasoners were presented with a description of a controversy and two alternate perspectives. For example, on one problem participants were informed that there may be a relationship between food additives and health outcomes, and that some research showed that food additives are unhealthy while other research showed that food additives make food safer to eat. The interview aimed to elicit the participants’ views of knowledge and concepts of justification via probes exploring participants’ opinions on the controversy, their explanation for their belief, their views on the intransience of the controversy, and explanations for how people can disagree about the controversy. As we reported previously, King and Kitchener (2002) concluded, based on responses to such vignettes, that few students demonstrated reflective thinking (or sophisticated epistemic beliefs). One strength of this measure is that it allowed participants to reason about scenarios. However, these scenarios did not provide evidence for participants to consider. Thus, the vignettes did not provide a context beyond the general topic. Further, King and Kitchener did not report on any of the strategies that epistemologically sophisticated students used to reconcile these problems—this suggested that students did not have knowledge of appropriate reasoning strategies. However, given how knowledge lean the problems were, it is not clear which reasoning strategies a non-expert could apply. A crucial question concerns how students’ might respond if they were provided with more information on the scenarios.
Kuhn and colleagues (2000) also examined students’ reasoning in order to assess their epistemic beliefs. They provided participants with two contrasting claims on topics of personal taste, aesthetics, values, and facts about the social and physical world. For example, two claims about the physical world that participants considered were, “Robin believes one book’s explanation of what atoms are made of,” and “Chris believes another book’s explanation of what atoms are made up of.” Participants were probed about whether they believed one or both claims could be correct. As we reported previously, Kuhn et al. (2000) concluded that students’ responses did not reflect the characteristics of evaluative thinking (sophisticated epistemic beliefs). Like King and Kitchener’s assessment, this activity aimed to allow participants to engage in reasoning about the scenarios. However, the alternate positions were presented in a way unconnected to actual evidence related to the controversies. How would students’ responses change if they were provided with just a bit more information, such as information that one book was a textbook and the other a science fiction book? Might students’ responses appear more evaluative?

The existing literature makes strong claims about reasoners’ EC, suggesting that it is unsophisticated and impoverished. However, we suspect that the tools used to evaluate EC are insufficient. The majority of research is based upon context-free, abstract assessments. Even interview studies that have provided situations to reason about require reasoners to draw extensively on existing background knowledge to generate evidence. Thus, our study contributes to the literature by providing students with a context before evaluating the sophistication of their EC. Why might providing evidence to respondents affect responses to such problems? We consider several possibilities.
One possibility is that problems with evidence can make it difficult to blithely offer a response before thinking it through. When presented with well-designed studies, it becomes harder to assert with little reflection that the studies on one side must be in error because reasoners must carefully think about the details.

Another possibility is that without evidence, reasoners might assume that it is difficult or impossible for anyone, including scientists or other experts, to generate reliable or valid evidence that bears on a problem. This arises from a lack of domain knowledge that specifies the kinds of data that are gathered in a field. When reasoners see actual studies, they can see not only that data can be brought to bear on these questions, but also what kinds of data can be gathered. This provides them with critical domain knowledge that might alter their judgments. For example, learners may argue that scientists may never be certain about dinosaur metabolism or extinction because dinosaurs are extinct. They may be unfamiliar with multiple forms of indirect evidence that they might find quite compelling if they were aware of it.

Third, high-level reasoning about multiple theories requires marshaling, evaluating, and weighing the strength of many pieces of evidence. Non-expert reasoners may need to have specific evidence provided to them because they lack the cognitive resources to simultaneously generate, track, and reason about theories, evidence, claims, and their own beliefs on their own. Thus, evidence could help reduce cognitive load and thus support the use of higher-level epistemic resources.

Fourth, it could be the case that in the absence of evidence, vignettes describing contrary theories might foster a Gricean assumption that the conflicts cannot be resolved. According to the Gricean principle of informativeness (Grice, 1989), a communicator
should provide the needed information to make a decision. If reasoners assume that this conversational norm is in play, they may assume that when it is simply asserted that there are studies that support each position, the principle of informativeness may imply that conflicting theories must be irreconcilable or intractable, or else the vignettes would provide information to make a reasoned response.

Finally, evidence might help reasoners pin down the meanings of the terms, ruling out certain kinds of responses. For example, King and Kitchener’s (2002) food additives problem (discussed above) seems to demand a general response about food additives. There is probably no general statement about the safety of all food additives that is possible. The presence of evidence can show that particular additives are being referred to; this allows reasoners to better apply epistemic resources to respond.

We suspect that existing measures of epistemic beliefs and beliefs about the nature of science do not provide students with enough contextual information to fairly evaluate their epistemic beliefs. Given that few people have extensive background knowledge in any particular domain, it should not be surprising that learners fail to generate, explain, and describe more complex theories, evidence, beliefs, and contextual relationships—that is demonstrate evaluativist EC. If reasoners’ responses are systematically different when they are provided with evidence, in addition to conflicting claims, this suggests that existing measures of EC fail to assess students’ true epistemological capacities. Rather, existing measures capture students’ heuristic, shallow thinking about knowledge.

Another complimentary approach to assessing EC is to evaluate students’ reasoning during learning activities. Mason and colleagues (2010) evaluated think-aloud
protocols during an internet search task for instances in which undergraduates explicitly reflected on the justification, source, certainty, and complexity of knowledge. The results indicated that many students reflected on evidence and the source of information, which suggests that students’ had sophisticated beliefs about the nature of knowing. On the other hand, only a few students’ verbalized their ideas about the certainty and complexity of knowledge during the search task. Mason and colleagues argued that the students had not developed sophisticated beliefs about the nature of knowledge. However, students were not explicitly prompted during this activity—it’s not clear that utterances about certainty and complexity would come naturally. We now consider how students’ patterns of reasoning may provide insights into their understanding of the certainty and complexity of knowledge.

Patterns of Reasoning

Although most researchers characterize the ability to engage in evidence-based reasoning as a key feature of sophisticated epistemic cognition (e.g. Chinn et al., 2011; Hennessey et al., 2013; King & Kitchener, 2004; Kuhn et al., 2000; Mason et al., 2010; Pluta et al., 2011), there is minimal research exploring this link. We aim to address this gap by utilizing students’ patterns of reasoning to understand their underlying epistemic beliefs. We propose that the best way to identify sophisticated beliefs is by examining patterns of reasoning across problems and by examining how providing students’ with evidence provides a context for engaging epistemic beliefs. We believe the EC literatures’ focus on the specificity of epistemic beliefs is misframed. Nuanced, highly specified, and adaptive beliefs mark sophisticated epistemic cognizers, in other words, evaluativists. It is critical to examine how reasoning changes across different contexts.
Learners’ beliefs about the nature of knowing are reflected in the targets of their reasoning. For example, epistemologically sophisticated learners believe that knowledge is justified by evidence, and therefore these learners are more likely to search for or request evidence, as well as concentrate their reasoning on the details of the evidence. Less sophisticated reasoners believe that knowledge is justified by personal opinion or that no justification is needed at all. Therefore, even if evidence is provided, these learners will discount or blithely ignore it during reasoning activities, choosing to target their reasoning on other task components. Similarly, reasoners who hold sophisticated beliefs about the source of knowledge will explicitly evaluate or search for source information, while epistemologically unsophisticated reasoners are more likely to ignore this information. As we have discussed, using the targets of reasoning to evaluate learner’s epistemic beliefs has been adopted by a number of researchers (e.g. Hofer 2004, King & Kitchener, 2004; Mason et al., 2010).

Learners’ beliefs about the nature of knowledge are often manifested in their patterns of reasoning across different types of problems and contexts. Mason and colleagues (2010) found that only a few students’ verbalized their beliefs about the nature of knowledge. We propose that a potentially good way to tap these beliefs during learning activities is by evaluating reasoning across multiple types of problems. For example, when reasoners believe that knowledge is complex, they will exhibit an array of reasoning strategies in order to develop an understanding of the complexity; when reasoners believe that knowledge is simple, reasoners will not see the need to apply multiple reasoning strategies to further understand problems. Epistemologically sophisticated reasoners should also recognize that complexity varies in different
dimensions and across different problems. For example, learners might recognize that some problems are complex because the focus phenomena are complex (e.g. a learner might believe that cellular processes are complex because of multiple, interrelated mechanisms), or that collecting good, useful evidence is where the complexity of certain problems exists (e.g. evaluating and characterizing dinosaur metabolism may be relatively straightforward. However, collecting good evidence bearing on this type of problem may prove challenging and be the source of problem complexity). These beliefs are reflected in the different types of reasoning strategies learners apply across problems. Reasoners learning about cellular processes might target their attention on the relationship between different mechanisms, while reasoners learning about dinosaur metabolism may target their reasoning on the quality of the evidence. Reasoners with sophisticated beliefs about knowledge should adaptively apply multiple strategies across problems, such as evaluating the quality of evidence, seeking new evidence, and generating alternate explanations. Reasoners with less sophisticated beliefs should demonstrate less or no variability in their strategy use.

Reasoners’ beliefs about the certainty of knowledge are also manifested in their reasoning across different types of problems and topics. Epistemologically sophisticated reasoners believe that the certainty of knowledge is content and context dependent (cf. Mason et al., 2010). Thus sophisticated reasoners’ patterns of reasoning should vary across different contexts. If learners approach a problem concerning the function of the stomach or description of specific bird species’ migration routes (relatively certain knowledge) in a different way than a problem concerning the speed of neutrino particles or explanation for how birds navigate utilizing magnetic fields (relatively uncertain
knowledge), it indicates sophisticated beliefs about the certainty of knowledge. A sophisticated reasoner may believe that the foundation of evolutionary biology is relatively certain, while details of specific evolutionary mechanisms are far less certain. Sophisticated reasoners are likely to utilize more reasoning strategies on less certain problems and fewer strategies on more certain problems. Thus, in discussing global evolution theory, a sophisticated reasoner might simply appeal to the authority of the vast majority of scientists who believe the evidence supports evolutionary theory, while in considering specific evolutionary mechanisms the reasoner may be more likely to see the need to elaborate on multiple perspectives, consider the evidence supporting or conflicting with the mechanisms, and evaluate the overall quality of evidence.

**Students’ Reasoning Ability**

The majority of studies exploring EC and beliefs about the NOS have not described students’ actual reasoning—despite most the aim of tapping students’ beliefs about evidence-based reasoning (King & Kitchener, 2004; Kuhn et al., 2000). It is reasonable to infer from this lack of description that that students’ simply did not engage in reasoning.

The research focusing on reasoning (rather than epistemic beliefs) is also not particularly helpful for assessing the targets and patterns of student reasoning. Cognitive scientists have primarily focused on describing students’ ability to apply singular reasoning strategies, such as, students’ ability to (a) evaluate experimental controls, (b) identify sources of bias and (c) evaluate the source of evidence.

Outside of a few case studies (e.g. Roth & Bowen, 1995), there are few studies that systematically evaluate students’ adaptive application of multiple strategies while
completing complex inquiry tasks (cf. Chinn & Malhotra, 2002). For example, research examining students’ evaluation of the source of evidence focuses on students’ ability retain information from credible sources and dismiss information from unreliable sources. Few, studies examine students’ ability to reason about multiple studies that include both (ostensibly) credible, albeit conflicting sources of information, as well as dubious studies. A second issue is that research on reasoning often focuses on expert-level application of strategies. For example, research examining students’ ability to evaluate experimental controls focuses on their capacity to systematically control variables during experimentation tasks. A general conclusion is that students’ have nascent knowledge of these types of strategies but without extensive instruction or practice, students’ lack expert knowledge of strategy application (cf. Zimmerman, 2007). For the evidence evaluation tasks commonly faced by laypeople, such as evaluating studies in news briefs, or tasks used to evaluate epistemic beliefs, strategy awareness may suffice.

Overall, there are few studies linking epistemic beliefs to targets and patterns of reasoning—despite the fact that sophisticated epistemic cognition is characterized as an appreciation and ability to engage in evidence-based reasoning (Allchin, 2012; King & Kitchener, 2004; Kuhn et al., 2000). In the present study we evaluate the extent to which this is a fruitful way evaluate the sophistication of learners’ epistemic beliefs. The existing EC research has not provided a careful description of the reasoning strategies middle school students’ have for coordinating evidence. Given the dearth of description, our working prediction is that students’ do not have the strategies needed to demonstrate sophisticated EC.
Overview and Goals of the Study

In this study, we examine the reasoning strategies seventh graders use to coordinate conflicting pairs of science study briefs. Students were presented with pairs of reports for up to four different science topics. The study reports were designed to be similar to the descriptions of studies found in newspaper articles. We aim to (a) describe students’ strategies for coordinating evidence, and (b) utilize patterns and targets of strategy to evaluate the sophistication and specificity of students’ epistemic cognition.

We address four predictions based on the results and assumptions of the dominant epistemic cognition literature:

Prediction 1. Students will have few, if any, strategies for coordinating scientific evidence.

Prediction 2. Students’ patterns and targets of reasoning will not be consistent with sophisticated epistemic beliefs.

Prediction 3. Providing students with contextual information, such as evidence, will have minimal impact on their reasoning and epistemic cognition.

Prediction 4. Students will demonstrate domain-level epistemic beliefs.

If these four predictions are supported, the dominant descriptions of EC and NOS in students’ remain accurate. However, if the predictions are not supported, it suggests that dominant descriptions are overly simple and ignore important aspects of epistemic cognition. More sophisticated measures will reveal more sophisticated beliefs! Some educational epistemologists have argued that very few people, even university graduates, attain the highest levels of epistemological development (King & Kitchener, 2002; Kuhn
et al., 2000); according to these researchers the highest level is when a person is an evaluativist, one who can carefully scrutinize evidence and claims in order to make the best possible judgments about verisimilitude of explanations and models (King & Kitchener, 2002; Kuhn & Weinstock, 2002). Conversely, it could be that simply giving people evidence to consider spurs an evaluativist state of mind. It might not be the case that people have impoverished beliefs about knowledge and knowing, but that their ability to bring up complex cases is inhibited, as well as their knowledge of reflective reasoning strategies.

Methods

Participants

Participants were 29 seventh grade students from a suburban middle school in the mid-Atlantic region of the United States. Within the school, 1% of students qualified for free or reduced-price lunch; 97% of students were white. Approximately 90% of students in the school reached proficient or advanced proficient levels on state language arts and mathematics exams.

Materials

For four scientific problems, handouts describing the problems and evidence were developed. The first section of each handout summarized the problem and briefly presented the two competing theories (33-122 words). Following this section were two brief research reports (99-149 words). Each report featured (a) a description of evidence for one of the theories, (b) an explanation of the evidence (c) a brief description of the data collection method, and (d) a final conclusion supporting one of the theories. The brief reports were designed to be somewhat similar to an abstract (See Appendix A for an example).
The topics of the reports were: (a) What causes frog deformities (Linn, Shear, Bell, & Slotta, 1999)? (b) Does aspartame cause cancer in lab mice? (c) Were dinosaurs warm or cold-blooded? and (d) Does poor diet or bacteria cause the disease pellagra? The key features underlying the coupled pairs were, respectively: (a) an interaction between causes, (b) failure to replicate, (c) indeterminacy due to lack of conclusive evidence, and (d) study falsification coupled with an argument from ignorance. The descriptions were brief enough that subjects could make strong cases for methodological flaws or inconsistencies. Topics were selected because students likely would not have a strong existing belief about relationships discussed in each topic (we referred to aspartame as AP-12 in student version of the studies, so not to engage students’ existing beliefs or concerns). In the remainder of this paper, we will refer to the problems as Frogs, Aspartame, Dinosaurs, and Pellagra. Table 3.1 describes the study pair problems.

**Table 3.1: Summary of Problems**

**Frogs**
Students read evidence that examined whether deformities in frogs were caused by pesticides or by trematodes (parasitic worms). One study was a laboratory experiment in which tadpoles and pesticides were placed in aquariums; then trematodes were placed in half of the aquariums. The tadpoles in the trematode tanks developed deformities and thus the scientists concluded that trematodes cause deformities. In the second study, scientists counted the number of deformed frogs found in natural ponds. Half of the ponds had a large amount of pesticides in them from farm run-off; the other half did not. Because only frogs in the ponds with detectable pesticides developed deformities, these scientists concluded that pesticides cause deformities. We were interested in whether students would recognize that there could be an interaction between the proposed causes.

**Dinosaurs**
Students were asked to consider whether dinosaurs were endothermic or exothermic. Some scientists suggested that because dinosaur fossils are found in cold regions, dinosaurs must have been warm-blooded to produce their own heat. In the second study, scientists had collected fossil evidence that dinosaurs had small pineal systems (brain structure responsible for regulating body temperature), evidence which suggests that dinosaurs must have been cold-blooded. We were interested in whether students would recognize that there was a strong element of indeterminacy in the problem.

**Pellagra**
Students considered evidence that suggested causes of the disease pellagra (this disease is
caused by a niacin deficiency). In one study, the number of outbreaks fell to zero after a sewer system was installed; scientists concluded that pellagra was caused bacteria spread by unsanitary conditions. In the second study, sanitary conditions were also improved in an area, and the rate of bacteria borne diseases dropped. However, cases of pellagra remained stable, and then increased as poverty grew in the study area. Scientist concluded that bacteria could not be the cause, and therefore poor nutrition must be the cause of the disease. We were interested in whether students would recognize that the second study served to falsify the initial study. Students could also comment on the extent to which correlation implies causation.

**Aspartame**

Students encountered studies in which the chemical AP-12 (Aspartame) was hypothesized to cause cancer in rats. Both studies were described such that the methods in each study were identical (this was even pointed out in the text of the second study). Despite identical methods, one study found higher occurrences of cancer in the rats. That study concluded that AP-12 causes cancer; the other study concluded that AP-12 does not cause cancer. We were interested in how students would deal with failure to of replicate.

**Procedure**

Each student was individually interviewed for approximately 40 minutes. Students read a single pair of studies and were interviewed on that topic before moving on to a new topic. Using a Latin square design, four different orders of the studies were formulated. Most students only had time to read and respond to three of the four pairs of studies. Nine students only had time to respond to the full set of questions for two problems. All interviews were audiotaped and transcribed verbatim. Our analysis focuses on students’ responses to the six interview questions listed in Table 3.2. Our questions were based on questions developed by King and Kitchener (2002) and designed to capture students’ beliefs about the nature of knowing and knowledge within the four domains. Interviewers were instructed to prompt students with follow-up questions including: “Do you have any more ideas?,” “Why do you think that?,” and “Explain more about that.”
Table 3.2: Interview Questions

(QA) How is it possible that scientists could conduct these two studies and get two different conclusions?

(QB) Which of these ideas is closest to your idea about these two studies and the scientists’ conclusions? [students selected from a list of responses which included these categories:
1. Probably neither conclusion is right.
2. Only one conclusion is right, and I’m pretty sure which one is right.
3. Only one conclusion is right, but I don’t know which one is right.
4. Both conclusions are right.
5. Both conclusions are partly right.
6. Both conclusions can be right, but one is probably better than the other.]

(QC) Is it possible to come up with an explanation that can make sense of both of these studies at the same time; what is it?

(QD) Do you think scientists will agree with your explanation?

(QE) Can scientists ever be certain about [the causes or relationships described in the studies]?

(QF) What actions should scientists take next to try to understand [the causes or relationships described in the studies]?

Coding Procedure

Coding schemes were developed for each of the six questions. The coding schemes are described in detail throughout the results section. In general, for each question an initial set of possible categories were developed based on the (a) scientific reasoning literature, (b) epistemic cognition literature, and (c) features of the problems themselves. Then, the first author conducted a card sort procedure with a subset of responses to develop and define the categories. Finally, the first author and a second rater coded subsets of responses individually and then compared ratings in order to further refine the categories and train the second rater. Once the coding schemes were finalized, both raters coded 25% of the data that had been set aside. For each of the schemes described below, interrater agreement of 85% or greater was obtained on the untrained data. Then two raters coded the full corpus, and discrepancies were resolved through discussion.
Results

We first consider strategy use, including strategy existence, variability, and the extent to which strategy application is tied to specific problems. We then consider students’ confidence in their own responses, their views on whether scientists will ever be certain, and their ideas about the next steps for reconciling the conflict. We conclude by evaluating their responses to each of the four problems: Frogs, Aspartame, Dinosaurs, and Pellagra.

Strategy Use

Our initial analysis concentrates on two interview questions that prompted students to critically evaluate and coordinate evidence. We jointly coded students’ explanations for why scientists came to different conclusions (QA) and their explanations aimed at coordinating the results of both studies (QC).

Strategy Existence. What reasoning strategies do students have for reconciling conflicting evidence? Evaluativist thinkers use evidence and reason to support their judgments.

We identified ten unique strategies for reconciling conflicting evidence. Table 3.3 presents descriptions and examples of each strategy. Our count somewhat underestimates the full range of strategies because categories were collapsed in a number of instances. For example, we found that students proposed at least three types of complex explanations. Complex explanations are instances in which students’ proposed an explanation that took into account the conclusions of both studies. These often included interactions, multiple causes, and mediating relationships. Because only a small number of students generated each of these more specific strategies, it was impossible to train coders with real examples and test the utility of a more fine-grained categorization
scheme. We describe some of less frequently engaged strategies when we consider students’ responses to each of the four problems in detail.

Three of the 10 strategies focused on discounting one of the studies. Students noted that studies could be discounted because of researcher bias (14% of students engaged this strategy for QA and 1 student engaged this strategy for QC) or methodological flaws (38% of students for QA, 31% for QC). A third strategy was to argue that one of the study’s conclusions did not follow from the data (coded as conclusion, 14% for QA, 7% for QC). Students who engaged these three strategies typically dismissed one study, allowing them to adopt the other study’s explanation. Overall, 66% of students utilized one or more of these three strategies.

Four of the strategies allowed students to preserve the conclusions of both studies. Some students identified methodological differences between the studies. We differentiated between responses focused on the design features of the studies (coded as design differences, 83% of students for QA, 24% for QC); responses focused on differences between the subjects of the studies (coded as subtyping subjects, 62% for QA, 48% for QC); and responses in which students noted that the studies were different without articulating any specific features, and further noted that the studies could not be compared (coded as incommensurable, 14% for QA, 10% for QC). An example of an application of the design difference strategy was when students proposed that the conflicting conclusions could be due to differences in the doses of aspartame used in each of the aspartame studies. An example of the subtyping subjects strategy was when students noted that the conflicting results could be due to different types of rats being used in each of the aspartame studies. Complex explanations were the fourth type of
strategy in this category of responses (52% for QA, 76% for QC). Students who engaged the complex explanation strategy used argued that information from both studies could to develop a new or more complex explanation of the phenomena. These four strategies allowed students to preserve the results of both studies, typically leading to a more complex explanation of the phenomena. All 29 students utilized one or more of these four strategies.

Two strategies did not seem to represent formal attempts to reconcile the conflicting studies—rather students’ responses pointed to a need for additional information. Some students responded that more information was needed to reconcile the conflicting conclusions (coded as abeyance, 10% for QC), this included calls for additional evidence or replication of the existing studies. Other students responded that the conflicting results were due to chance (14% for QA, 10% for QC). Overall, 28% of students utilized at least one of these two strategies.

Finally, some students failed to develop cogent explanations for Questions A and C (coded as other response, 17% for QA and 41% for QC). These types of responses ranged from summarizing the problem or evidence, to saying “I don’t know,” or responding in ways not well-aligned with the questions. Due to the novelty, difficulty, and time constraints of the task, as well as fact that no students used this strategy consistently across all problems they encountered—we judged that this was actually a reasonable approach. Sometimes scientists just don’t know the answer—so a response such as, “I don’t know” is fair. Summarizing or responding in way not well-aligned to questions may represent attempts to puzzle through the problems.
The right hand columns of Table 3.4 and 3.5 summarize the number of instances of strategy engagement aggregated across all problem types for Question A and Question C respectively. Overall, the four most common strategies used to explain why scientists came to different conclusions (QA) and when generating explanations to fit both studies (QB) included (a) proposing complex explanations, (b) identifying methodological flaws, (c) appealing to design differences, and (d) subtyping subjects. This was true for both explanatory questions and these four strategies were used across all four topics. Overall students collectively demonstrated awareness of a nice arsenal of evidence-focused reasoning strategies.

Table 3.3: Strategies and Brief Description

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Brief description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method Flaw</td>
<td>One or both of the studies had a mistake, flaw, or confound (explicitly stated, otherwise the responses was coded as “method difference”).</td>
<td>It’s possible because something could have gone wrong in one of them, like even though they tried to clean out all the viruses or bacteria, still there could have been some that did cause it. Or like, maybe people were still unhealthy—like people who they thought had poor diets, they were probably living in a place where there’s a lot of bacteria and bad stuff that caused it. So it could have been a mistake that nobody really realized. (Pellagra)</td>
</tr>
<tr>
<td>Design Difference</td>
<td>Different results were attained because researchers used different designs or measures. Includes features of the study such as the location/environment or independent variables. It does not include subjects.</td>
<td>It doesn’t say what the different amounts are. So what could be very high for one, for the other could have been a medium amount. We don’t know the exact amounts. (Aspartame)</td>
</tr>
<tr>
<td>Complex Explanation</td>
<td>Study designs/conclusions might ignore an interaction/relationship, multiple causes/relationship, or other underlying causes/relationships.</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>One or both of the conclusions are dominated by experimenter bias, subjectivity, or opinion.</td>
<td></td>
</tr>
<tr>
<td>Chance</td>
<td>Differences in results are due to chance/randomness.</td>
<td></td>
</tr>
<tr>
<td>Conclusion</td>
<td>Conclusions/explanations do not follow directly from the evidence (this might also include correlational relationships).</td>
<td></td>
</tr>
<tr>
<td>Subtyping Subjects</td>
<td>Student explicitly says that the subjects of the study were different (frog, dinosaur, rats, citizens).</td>
<td></td>
</tr>
</tbody>
</table>

Well, if this one... maybe perhaps the pesticides attract to trematodes... (Frogs)

This group got paid off... There was a lot of incentive to put these additives and drugs forward and theoretically it would be very hard to do, but you could bribe these scientists into altering their experiments so it comes out like that. (Aspartame)

Well it probably happens randomly within each other, because—well, I believe that the trematode is the cause. So I think it’s random. There could be—the trematodes could be in the pesticides. (Frogs)

They didn’t really know for a fact that the people couldn’t get good food.... Because in study two they just assumed that it came from a poor diet because most people got it probably couldn’t get good food. (Pellagra)

Well, they could search like up north—like in Canada or something like that. And it’s colder up there, so they could have found cold blooded animals, or dinosaurs like that, the duckbill dinosaur. And then if they went down towards the equator, they might find fossils of warm blooded animals. So they could probably find different things for each dinosaur. (Dinosaurs)
Incommensurable

Different results were attained because researchers used incommensurable methods or measures. Student made absolutely no attempt at reconciling the results because she believes it is impossible to do so.

Well, here they’re doing the pineal system of the body, and it’s basically saying that it controls the body temperature, which is like saying, “Ok, you’re body’s warm. But where you’re living is cold.” So if you’re living somewhere cold, you must be warm blooded or else you would have died. And if study 2 is saying that you’re studying the body and the body’s making it warmer, then they must have been warm blooded. It’s like taking two different variables and trying to make them one variable. You can’t combine two different variables as one. Like X and Y. You can’t make them both X.

(Dinosaurs)

Abeyance

Defers explanation, awaiting more information (typically suggests that there is an explanation that could account for both studies).

I’m not sure that this study is right. I think that they could do it over and try to do the same, both outside of the body or both inside of the body...I think it is possible but I think they should do another experiment to try... I don’t really know. (Dinosaurs)

No Response

No attempt or failed attempt to answer the question (includes summaries, when students answer a “different” question and “I don’t know” responses).

I think it's possible, I just can't think of a logical reason right now. Trying to think of one but...I just can't think of it...I think if I had a little more education on this topic it definitely would be possible for me to think of one but winging it off of here, I don't think I could think of anything. (Dinosaurs, Student 7)

**Strategy Variability.** To what extent do students have the knowledge, access, and ability to use strategies adaptively? If individuals only used a single strategy to reconcile all encountered problems, it suggests that middle school students struggle to
recognize and grapple with the differences in scientific problem types. Using different strategies for different problems suggests that students are attentive to important contextual details when engaging in scientific problem solving—a feature of evaluativist thinkers and topic-specific epistemic beliefs.

When students were asked to explain how scientists could generate different conclusions (QA), all 29 students applied more than one strategy across the different problems they encountered. When we asked students to generate an explanation to fit both studies (QC), 97% of students used more than one strategy across the different problems they encountered. Across all problems and both questions, we found that students demonstrated an awareness of an average of 4.2 reasoning strategies (range: 2-6). Students clearly demonstrated variability in their scientific thinking, even with our relatively coarse-grained scheme.

**Sensitivity to Contextual Information.** To what extent is strategy use aligned with specific problems? If patterns of strategy application emerge, it suggests that students may be able to reflectively coordinate strategies to specific problems.

Tables 4 and 5 display the number of students who engaged different strategies in response to the four problems for Questions A and C. For the question in which students were instructed to explain how scientists could generate different conclusions (QA), students primarily appealed to general design differences for the four problems. Problem-specific responses were more apparent in the second most dominant strategy applied to each problem. For example, for the Aspartame and Dinosaur problems, approximately one third of students noted that the conflicting results could be due to different types of subjects (subtyping subjects) being used in different problems (38% and 32% of students,
respectively). In comparison, no student subtyped subjects for the Frogs problem and only two (7%) subtyped subjects for the Pellagra problem. For the Pellagra problem, the second most dominant strategy was to propose a complex explanation (38%). For the other problems, the proposal rate for complex explanations ranged from 27% (Dinosaurs) to 13% (Aspartame). For the Frogs problem, the second most dominant strategy was to argue that one of the studies had a methodological flaw (23%).

When students generated their own explanations to account for both studies (QC), their preference for considering different kinds of subjects (subtyping subjects) in their explanations remained for the Aspartame (29%) and Dinosaur (32%) problems. In comparison, no students proposed a subtyping subjects explanation for the Frogs and Pellagra problems. Instead, students’ responses shifted to explicating complex explanations for both the Frogs (59%) and Pellagra (56%) problems.

Overall, Tables 4.4 and 4.5 indicate that the middle school students’ strategies showed some alignment with specific problem types.

**Table 3.4: Count data across problem types for Question A (QA), “How could scientists come up with two different conclusions?”**

<table>
<thead>
<tr>
<th></th>
<th>Aspartame N=24</th>
<th>Dinosaurs N=22</th>
<th>Frogs N=22</th>
<th>Pellagra N=27</th>
<th>Overall N=29</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design Difference</td>
<td>10</td>
<td>6</td>
<td>14</td>
<td>12</td>
<td>24 (83%)</td>
</tr>
<tr>
<td>Complex Explanation</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>11</td>
<td>15 (52%)</td>
</tr>
<tr>
<td>Subtype Subjects</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>18 (62%)</td>
</tr>
<tr>
<td>Method Flaw</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>11 (38%)</td>
</tr>
<tr>
<td>Conclusion</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4 (14%)</td>
</tr>
<tr>
<td>Incommensurable</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4 (14%)</td>
</tr>
<tr>
<td>Chance</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4 (14%)</td>
</tr>
<tr>
<td>Bias</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>4 (14%)</td>
</tr>
<tr>
<td>Abeyance</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>No response</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5 (17%)</td>
</tr>
</tbody>
</table>

*Students could utilize more than one strategy per problem.*
Table 3.5: Count data across problem types for the Question B (QB), “Can you come up with an explanation that fits both studies?”

<table>
<thead>
<tr>
<th></th>
<th>Aspartame N=24</th>
<th>Dinosaurs N=22</th>
<th>Frogs N=22</th>
<th>Pellagra N=27</th>
<th>Overall N=29</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design Difference</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>7 (24%)</td>
</tr>
<tr>
<td>Complex Explanation</td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>15</td>
<td>22 (76%)</td>
</tr>
<tr>
<td>Subtype Subjects</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>14 (48%)</td>
</tr>
<tr>
<td>Method Flaw</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>7 (24%)</td>
</tr>
<tr>
<td>Conclusion</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2 (7%)</td>
</tr>
<tr>
<td>Incommensurable</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3 (10%)</td>
</tr>
<tr>
<td>Chance</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3 (10%)</td>
</tr>
<tr>
<td>Bias</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>Abeyance</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3 (10%)</td>
</tr>
<tr>
<td>No response</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>12 (41%)</td>
</tr>
</tbody>
</table>

*Students could utilize more than one strategy per problem.

Theory Choice Variability

We queried students about which of the scientists’ conclusions they preferred for each of the four problems (QB). Students selected from six choices provided on a sheet of paper: (1) probably neither conclusion is right; (2) only one conclusion is right, and I’m pretty sure I know which one is right; (3) only one conclusion is right, but I don’t know which one is right; (4) both conclusions are right; (5) both conclusions are partly right; and (6) both conclusions can be right, but one is probably better than the other.

Seventeen of 29 students (59%) responded in three or four different ways across the problems they encountered. The remaining twelve (41%) students responded in two different ways. Every student demonstrated variability in their responses across problems. This result provides additional evidence that students were carefully weighing the evidence that they had access to for each problem type; this is indicative of sophisticated epistemic cognition.

Confidence in Own Explanations
To what extent are reasoners confident in their own explanations accounting for the evidence? We examined whether students believed if scientists would agree with students’ own explanations of the phenomena (QD). Reasoners’ confidence in their responses to ill-structured problems may be a good proxy for their beliefs about the complexity of the problems. Reasoners’ who say that scientists would agree with their conclusion may believe that the problem is relatively simple—so simple that even middle school students can generate reasonable explanations.

We organized students’ responses into four categories. Across all four problems, 59% of students responded that scientists would agree with students’ own explanation for at least one of the problems they encountered. Thirty-one percent of students argued that scientists would disagree with students’ own explanation. Twenty-four percent of students explicitly noted that some scientists would agree and some would disagree. Fifty-two percent of students’ provided responses that suggested that they were unsure if scientists would agree or disagree with their explanations. Finally, we identified two instances (7% of students) in which the students seemed to have been answering a different question entirely—these responses are not considered further.

These responses suggest that many students see some scientific problems as simple. This result is consistent with existing research that view have described students’ EC as unsophisticated. We now consider the extent to which students’ responses varied across problems.

**Variability.** To what extent did students’ confidence in their own responses vary across problems? We found that 24% of students believed that scientists would agree with their explanation across all of the problems they responded to. One (3%) student
believed that scientists would not agree with his explanation for all problems he responded to. The remaining students (72%) provided a combination of different types of responses or only provided uncertain/indecisive responses.

Here is an example of a student whose evaluation of whether scientists would agree with his explanation changed across the three problems he encountered. In responding to the Aspartame problem, the student argued that different doses of aspartame may have been used by the scientists and that smaller doses may not cause cancer, while larger doses may. The student responded that scientists would not agree with that explanation:

No I don't think that because they each have their own opinion. And they probably just agree with their own studies. So they probably don't agree with each other. (Student 9)

This student’s response may suggest awareness that this type of scientific issue can be both scientifically complex, as well as socially charged (socially complex). For the Frog problem, this student argued that both trematodes and pesticides were both causes of deformities. He responded that scientists would agree with his explanation:

Yeah I think that they would because they would both see how both the trematodes and the pesticides, how they're both affecting the frogs, and how they can both be right. (Student 9)

It seems the student had confidence in the two studies and that scientists would readily adopt of his ideas. Why would this response be different from his response to the aspartame study? Perhaps the student sees the Frog problem as less scientifically and
socially complex. Unlike the Aspartame problem, scientists would be less likely to be driven by their preexisting “opinions” (or theories or biases) as in the Aspartame problem. For the Pellagra problem the student argued that the second study was flawed, and thus the conclusion of first study (bacteria as the cause of pellagra) was correct. He stated that some scientists would agree and others would not agree with this explanation:

I think Study 1 would agree with me, but maybe not Study 2 because they probably would have thought that they cleaned up everything, like they did everything that could get rid of the bacteria and viruses. And the reason why the amounts of pellagra kept on going up is a reason because of the poor diet. So I don't really think that the scientists that were a part of Study 2 would agree with me, but Study 1 would. (Student 9)

It seems that the student was (again) addressing the social complexity of science—scientists may be less likely to adopt the views of those who critique their scholarship. It is less clear from this response how complex the student views the scientific component of the problem to be. Across these three problems, we see this students responding to the contextual information in the problems in order to calibrate his responses.

As we noted above, students often did not unequivocally declare how scientists would respond:

Um, I don't know. They might not agree because they might think that they're right, when I don't know if they're right. But they also could go with my
explanation because they both could be right. Because potentially they could be studying the same dinosaur or the same place. But they also could not agree because they're sort of different-- they got two different things from two different places. So they could both be right and agree, or both could be wrong and not agree because they'd probably go with what they found (Student 6).

For the purposes of this study, we viewed uncertain/indecisive responses as indicators of sophisticated epistemic cognition because students who provided these types of responses seemed to be carefully considering multiple perspectives.

Overall, the majority of students demonstrated variability in their responses across multiple problem types that is indicative of sophisticated EC. Many students who did not show explicit variability did demonstrated careful perspective taking by providing uncertain/indecisive responses. However, compared to the previous two questions, a smaller number of students showed variability in their responses across problems. All but one these students declared that scientists would agree with their explanations across problems—which, we have argued, is evidence of holding the belief that scientific knowledge is simple.

**Reasoning.** What reasons did students commonly provide for their confidence in their response? When students argued that scientists would agree with their explanation, they typically elaborated on how their explanation fit the evidence. We also identified four common reasons that students provided for why they believed scientists would not agree or with their proposals. This scheme is summarized in Table 3.6.
Table 3.6: Responses to Question D (QD): Students’ Reasons for Why Scientists Would Not Agree With Their Explanation.

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Brief Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notes the role of evidence in determining whether scientist will accept students’ own explanation.</td>
<td>I think they would have to do more studying first, like for the scientist in study number 1 might want to do study with just trematodes, before they agreed with it.</td>
<td></td>
</tr>
<tr>
<td>Student Limitation</td>
<td>Student suggests that lack of background knowledge, experience, or credential limits the uptake her explanation.</td>
<td>I don’t know. I’m not a scientist. I don’t know anything.</td>
</tr>
<tr>
<td>Student Impoverished Explanation</td>
<td>Student notes that there are likely to be weaknesses or flaws in their explanation.</td>
<td>I need to have a little better explanation because scientists need details. I have no details, so they wouldn’t agree.</td>
</tr>
<tr>
<td>Scientists’ Bias or Disposition</td>
<td>Mentions that dispositions, biases, or opinions of scientists could influence uptake of ideas.</td>
<td>If they’re not too big-headed. I think an open-minded scientist would like to study all of this, not just their own explanation.</td>
</tr>
</tbody>
</table>

Across all problems, Fifty-nine percent of students explicitly reasoned that scientists’ acceptance of students’ own explanations was dependent on gaining access to additional evidence. Thirty-five percent (6 of 17) of these students provided this type of response more than once. The prevalence of this response suggests that many students see expert scientific inquiry as an extended process and requires the collection of a significant amount of evidence—this is consistent with normative epistemic beliefs about the nature of science.

Three other types of responses concerned potential deficiencies in student or expert thinking. Fourteen percent of students suggested that acceptance of their ideas would be unlikely due to their own cognitive limitations and credentials. Ten percent of students stated that they believed their own explanations were impoverished. Finally,
28% elaborated on biases or dispositions that scientists could have towards new ideas. Overall, 52% of students provided responses focuses on the cognitive dispositions of scientific thinkers (students’ or scientists) engaging in reasoning.

Here is an example of a student who demonstrated awareness of a number of these issues in a single response:

I don't know because to generate this sort of proof, it probably takes months and months, and quite honestly I think they'd be angry if someone with no knowledge of the subject came in and tried to disprove what they've already done. But, I don’t want to tell them that I'm right and they're wrong. Just, if I could, I'd love to throw some sort of idea with them, and even if they prove it wrong I'd love to know the right answer (Student 27).

This student recognized that: science is an extended process requiring evidence (“proof”), scientists have certain dispositions which may inhibit uptake, there are limitations in his own knowledge or credential that could influence the uptake of his ideas and the uniqueness of ideas may contribute to the scientific process. These types of responses are associated with sophisticated epistemic beliefs about the nature of science.

Overall, many students demonstrated evaluativist cognition when deciding whether scientists would agree with their responses. The majority of students provided different responses across problems or expressed uncertainty across all problems. As we have belabored, this variability is an indicator of sophisticated EC. Further, a significant number of students provided reasons for their responses that were consistent with
evaluative views of knowledge, such as noting the role of evidence, cognitive limitations, or expert bias. The balance between evidence and scientific thinkers’ dispositions suggests an understanding of both subjective and objective aspects of knowing.

**Certainty of Knowledge**

Do students’ evaluations of the certainty of knowledge vary across problem topics? After asking students to grapple with the study pairs, we invited students to judge whether scientists would ever be certain about the causes of the four phenomena (QE).

We found that in their open-ended responses students regularly hedged, making it difficult to evaluate their explicit views or beliefs about certainty. They frequently used words such as maybe, probably, possibly, might, not really, and probably not. These types or responses were coded as uncertain responses. We also coded responses which suggested that students concretely stated that scientists will never be certain and definitely be certain. Forty-six percent of students said that scientists will never be certain and 67% said that scientists will definitely be certain for at least one problem. Sixty-three percent provided at least one response that fell into an uncertain category.

**Variability in Certainty Judgments.** After having students consider only two pieces of evidence that conflicted, we found that individual students had many different ideas about whether scientists would ever be certain across problem types. Sixty-seven percent evaluated the certainty of topic knowledge differently across problems. This suggests that these students took into account a range of evidence, research methods, their evaluation of the complexity of the problem, and other contextual factors in order to make a judgment about the certainty of knowledge. Thirteen percent of students stated that they were uncertain if scientists would be certain across all problems they responded.
Seventeen percent claimed that scientists would definitely be certain across all of the problem types that they evaluated.

In order to illustrate how robust students’ sensitivity to context was, we consider examples of how students shifted their views about certainty across problems. Here of an example responding that scientists will never be certain for the Frogs problem:

This particular one, I don't think they can. I think that deformities can be caused from anything, not just trematodes and pesticides in the water. I think, like humans it could be, um, it could be genetic when the tadpole eggs are laid and I don't think they'll be able to I think that if they test the eggs and find that there's no deformities in it then um and then they see uh with both of the areas with pesticides and trematodes, I think then they'll have a better understanding. But not one complete answer (Student 7).

This student response suggests that she believes the problem is too complex for scientist to ever be completely certain. It is worth noting that this student utilized information about genetics cause that did not appear in the study, rather than working to coordinate study’s details. In the Dinosaur problem, this same student believed that one reason why scientists have not discovered that the nature of dinosaur metabolism is that evidence is scarce, not that the problem is particularly complicated:

Um, I think they definitely can be certain if they find the information that they need and I don't think it's definitely going to just pop up out of nowhere one day. It's really gonna take a lot of planning and good expedition, well you have
excavation for fossils, but I think they probably could come up with an explanation for it (Student 7).

Here is an example from a less loquacious student; he believed that scientists will never be certain for the Aspartame problem:

I don't think they can because it's like kind of confusing of how you can figure out the different-- how you could have the same experiment but it comes out as two different answers (Student 21).

Yet, in the case of the Pellagra studies, he believed that the problem is very simple:

Yeah, I think scientists can agree that both of them can cause it since both of them actually do make you have pellagra. (Student 21)

These two students’ responses illuminate the extent to which learners’ representations and strategies for coordinating the studies vary across domains and significantly influence their assessment of certainty.

**Reasoning Underlying Certainty Judgments.** We identified two common explanations in students’ judgments about whether scientists would ever be certain.

Forty-six percent of students referenced evaluating or testing variables other than those directly mentioned in the studies for at least one problem (complex problem space code). These students conceptualized the problem space larger than suggested by the studies alone. These students frequently indicated that they believed scientist would never be certain or that it would take significant time and resources for scientists to be certain. For example, for the Frog problem some students believed that for scientists to be certain, they would have to investigate every single variable that might influence or cause
deformities including every type of pesticide, parasite, pond water, pond size, frog species, and so on:

I think pesticides is an interesting idea, and I would think about it more if they had a more controlled study. But I think it-- you could have an absolute answer, but it would take a lot of tests with a lot of variables -- not only trematodes and pesticides, but all sorts of stuff to have an absolutely 100% right answer. It'd be hard. [Can you say maybe just a little bit more about that?] Well, you could say what's in the pond with the frogs. Maybe what else is contaminating the water, like some sort of waste if it's in the water. Like animal waste, even human waste, could be in the water and then maybe- (Student 27).

This type of response suggests that when producing certainty judgments, students frequently consider how many variables scientists may need to test and rule out, the accuracy of the methodology and measures, and even the number artifacts there are to test or study. Students who respond in this way see scientific problems as complex.

In the Dinosaur metabolism problem more students argued that scientists would never be certain, than in any other problem type. Seventy-nine percent of students who responded to the certainty question for the Dinosaur topic mentioned that it would be difficult for scientists to either collect enough data or trust the data because dinosaurs have been extinct. This indicates that students carefully process and utilize contextual cues when making judgments about certainty when problems have adequate context.

Overall, students demonstrated variability in their judgments about whether scientists will ever be certain about the scientific phenomena. Their responses seemed to be well-aligned with the contextual features of the studies. Researchers have argued that
when students say experts will never be certain about a problem it suggests that the students’ are multiplists or relativists (e.g. King & Kitchener, 2004; Kuhn et al., 2000). However, our results indicate that there is more behind this type of response—students may actually be signaling the complexity of problems or that it may be difficult to collect compelling, useful data. We think this is indicative of sophisticated EC.

**Science as Reasoned Inquiry**

What is the ensuing phase in the inquiry process? Our final question (QF) required students to propose new steps for scientists to take to resolve the conflicts. Responses provided information on students’ beliefs about the justification and complexity of scientific knowledge. If, for example, students propose that significantly more research is needed to resolve the conflicting results, it suggests that students see knowledge as justified by the quantity of evidence supporting it. Further this type of response may indicate that students see the problem as relatively complex because it requires additional inquiry to resolve the complexity. Other students may see knowledge as being justified by a single, well-designed study—this response may be associated with the belief that the topic is relatively simple.

We identified four main approaches to responding to this question: (a) describing a specific study (usually singular), (b) calling for more research (quantity of evidence), (c) calling for replication of existing research, and (d) suggesting that scientists communicate amongst themselves. In addition, 5% (4 of 78) responses did not fall into these categories and are considered separately.

Eighty-seven percent of students generated relatively specific studies to reconcile the conflict for at least one problem. For example, they proposed improvements to
existing designs, such examining a wider range of aspartame doses or examining pesticides in laboratory controlled settings. They also generated entirely different studies and techniques, such as utilizing “dissection” to identify traces of pesticides or parasites in deformed frogs or comparative studies for the dinosaur problem. Examples are provided in Table 3.7. Overall, there were a wide range of techniques and methodologies proposed. Most examples were not prevalent enough to reliably train coders. Previous research has documented that typically only specialized scientists can propose practical experiments—nonspecialized scientists propose logical and reasonable experiments, but which are deemed impractical by experts with more specialized knowledge (Shraagen, 1993; Schunn & Anderson, 1996). Our judgment is that most students generated reasonable variables to focus research on or new methodological approaches which were sensible given the limited amount of content and methodological knowledge they possessed. All responses of this type suggested a keen understanding of the role of evidence in inquiry.

Table 3.7. Responses to Questions F (QF): Examples of Students Specific Studies.

<table>
<thead>
<tr>
<th>Proposed Study</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved Controls</td>
<td>I think that some more lab controls things—I mean I honestly want to see what would happen in a lab-controlled environment with pesticides. (Frogs, Student 27)</td>
</tr>
<tr>
<td>More Differentiated</td>
<td>One experiment, mix the AP-12 into the same amount of food for the same batch, separate the rat groups so that you have one group that gets a small dose every day, one group that’s getting a medium dose, and one group that’s getting a large dose, instead of the same group getting different doses at different times. So that you would be able to determine which one causes the cancer and which one doesn’t. (Aspartame, Student 8)</td>
</tr>
<tr>
<td>Variables</td>
<td></td>
</tr>
<tr>
<td>Dissection</td>
<td>They should—the old fashioned way I guess. Dissect frogs. They could look for traces of pesticides and trematodes in them. (Frogs, Student 1)</td>
</tr>
<tr>
<td>Comparative Studies</td>
<td>I think that they should find other animals, kind of like</td>
</tr>
</tbody>
</table>
dinosaurs, maybe other animals that relate to dinosaurs kind of, and test the animals and see what they come out as — either warm blooded or cold blooded. And maybe that can lead to more information about the dinosaurs because the dinosaurs—other animals probably evolved from dinosaurs so they’re kind of in the same category. (Dinosaur, Student 25)

Specific New Studies

They could expose somebody to it and get somebody who has a good diet and expose it to them and see if they get it. And a person who has a poor diet and see which one of them gets it or something else like that. (Pellagra, Student 23)

Twenty-eight percent of students explicitly noted that more evidence was needed (quantity of evidence) for at least one problem and typically noted that the evidence should come from a diverse array of methodological approaches. Here is an example of a student who responded that the quantity of evidence required is aligned with the complexity of the problem and that it may require a significant and diverse range of evidence for scientists to converge on an explanation:

If they did this maybe a few years from now, they would probably have almost every test as possible completed. So I think they’re going to all lead up to two different major groups, only because they’re basically both right just they both have different explanations: whether it does cause cancer or whether it does not. And it’s just a very complicated topic because there’s so many different possibilities or tests that you could do in order to find the answer. So I think it’ll be a while before they actually take action and find out (Student 30).

This amount of elaboration was somewhat atypical. Most students did not explicitly tie the amount of evidence needed to the complexity of the problem:

They should just, like I said, conduct a wider variety of experiments and get more tests and controls and stuff and see how it goes (Student 5).
Overall, a significant number of students noted that the quantity and diversity of evidence is important to reconciling conflicting explanations.

Twenty-four percent of students proposed that the studies should be replicated for at least one problem. None proposed this action more than once. Responses were categorized in this way if students said that a study should be done again or over (no student used the specific word “replicated”).

Seventeen percent of students proposed that scientists should aim to communicate their results to other scientists. Two students responded in this way for more than one problem. We think this was particularly important response that demonstrated sophisticated EC: Dissemination of knowledge is central aim of epistemic communities (e.g. Goldman, 1999), but is not highlighted in most traditional instructional settings.

We now turn to the three types of responses that did not fall neatly into the four primary response categories (these were formally coded as other by raters). We found that one student argued that the experimental research should be suspended. For the Aspartame and Frogs problems, she expressed concern for the research subjects (she did not respond for the other two problems):

I think that they should stop giving it to any animal or test if it works on any other animal. Just make sure so they don’t keep on giving it to animals without knowing then it gives them cancer also. So just stop giving the rats the AP-12.

(Student 17)

Her responses fell outside of the areas traditionally emphasized by EC researchers.
Another student, when discussing the Dinosaur Metabolism problem, keenly argued that it was not clear what steps scientists could take:

I don’t think there’s any way that they can actually find out, because if there was they would have to go back in time to actually know what a dinosaur is. (Student 22)

Kuhn and colleagues’ or King and Kitchener’s EC frameworks may have categorized this type of response as being indicative of a multiplist or relativist thinker. However, when considering more contextual information, as well as the students’ overall pattern of reasoning (the students proposed a specific dissection study in response to the pellagra problem)—this student seems to be demonstrating evaluativist thinking—carefully considering problem-specific knowledge to generate a problem-specific strategy. Finally, one student simply stated he did not have a good idea of what next step scientists should take for the Frogs problem. As we have previously argued, we think that this type of response is quite reasonable for a sophisticated epistemic thinker—sometimes evaluativist thinking requires time.

We found that most students argued that a single critical study could be used to reconcile the conflicting results. Students had an array of ideas on how to do this. Many students’ also pointed to a need for a larger body of research to reconcile the conflicting evidence. Finally, we found that some students proposed that scientists simply communicate better to reconcile results. Once again, we see that if provided with an appropriate context to respond to, students demonstrate a rather sophisticated understanding of various aspects of the nature of science.

Responses to Specific Problem Types
In this section, we examine each of the four problems. Each problem had an underlying feature; we are interested in whether students would address these features.

**Frogs.** Biologists believe that frog deformities are caused when pesticides weaken frogs’ immune systems, which in turn increases the likelihood of parasites infection. The Frog problem presents students with one study that concluded that pesticides cause of frog deformities and a second study that concluded that parasites were the cause. The designs of these two studies did not rule out the possibility of an interaction. We were interested in whether students would infer a model that takes into account the details of both pieces of evidence and how closely the models would align with the casual explanation proposed by biologists.

We identified seven distinct types of causal explanations in student responses. Examples are described in Table 3.8. Students used six different reasoning strategies to generate these seven causal explanations. For example, one student said he believed that trematodes were the cause—chalking up the conflicting results to chance or randomness. The majority of students (59%, 13 of 22) proposed a complex explanation that took into account aspects of both studies. However, within that group, only two students were able to generate an interactive relationship between pesticides and parasites approaching biologists’ explanation. One of these student believed that the pesticides attracted the trematodes. The second student proposed that the trematodes needed the pesticides to cause deformities, but no explicit mechanism was stated. The majority of students argued that both pesticides and trematodes were causes of deformities and proposed that both were acting on the frogs. It was unclear if this type of response was due to light reading
of the studies or if students did not have enough knowledge of biological mechanisms needed to infer that there was a probable relationship between the causes.

Instructionally, if presented to groups of students, the corpus of seven explanations might spur more in-depth conversation about the link between the explanations and evidence. The diversity of ideas may support rich discussion that would allow students to converge on more normative explanations and flesh out the extent to which they are supported by the evidence.

**Table 3.8. Models Proposed by Students for Frog Deformity Problem.**

<table>
<thead>
<tr>
<th>Explanation/Models</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both pesticides and parasites cause frog deformities</td>
<td>What they are seeing that trematodes and pesticides both cause deformities, like trematodes might cause the other 70-80% and because pesticides what they found, in the second study is that pesticides cause 20-30% of the deformities so maybe, them together, like, they both, cause deformities, but, them separately, only cause part of the population to have deformities (Student 31).</td>
</tr>
<tr>
<td>Only parasites cause frog deformities</td>
<td>Well it probably happens randomly within each other, because-- well, I believe that the trematode is the cause. So I think it's random (Student 1).</td>
</tr>
<tr>
<td>Both pesticides and parasites cause frog deformities, depending on the kind of frog.</td>
<td>So I guess that maybe they could be a different kind of frog, and they could have-- I guess only the trematode could be attracted to that kind of frog, or for some reason only like that kind of frog and bury into it and feed off of it (Student 15).</td>
</tr>
<tr>
<td>Both pesticides and parasites cause frog deformities, depending on the environment.</td>
<td>Maybe in bigger ponds pesticides cause more frog deformities, and in smaller ponds trematodes cause more frog deformities (Student 5).</td>
</tr>
<tr>
<td>Pesticides and parasites interact to cause frog deformities.</td>
<td>Well, the trematodes, the study number one, maybe they need the pesticides to cause the deformities in the frog. Like cell respiration needs oxygen, all those things, oxygen, glucose to make the stuff it makes. So maybe the trematodes need the pesticides to deform the frogs. Because they did not check plain old trematodes. Everything has pesticides (Student 19).</td>
</tr>
<tr>
<td>Pesticides cause (attract) Parasites/Trematode which</td>
<td>What if the pesticides attract trematodes? In this one the... actually, maybe... nevermind that explanation</td>
</tr>
</tbody>
</table>
cause frog deformities. won't work though because if you had it here, there's no way that it could attract the trematode, like I said. So no, I don't think there's an explanation (Student 8).

Some other cause causes frog deformities. Because a parasite, unless it injects some sort of virus or something in, I'm not sure about it (Student 27).

Aspartame. Based on research reviews, government agencies such as the United States’ FDA and the European Food Safety Authority have deemed aspartame safe for consumption. However, this research body has been rife with conflicting results. Studies have been critiqued for having methodological problems, such as small sample sizes or models in which the laboratory mice were not of an appropriate age. Others have accused researchers of bias because manufacturers funded them. In the present study, students evaluated two studies with identical designs. However, the results did not replicate: one study concluded that aspartame caused cancer in rats and the other concluded that aspartame did not cause cancer. We were interested in what explanations students would generate for failures to replicate.

We identified three common strategies that students used to generate explanations that fit both studies. Students argued that (a) there were methodological flaws, (b) there were unreported design differences, or (c) researchers used different types of rats (subtyping subjects). Further a total of 11 (of 19, 58%) students explicitly noted that more research was needed. Each of these strategies was well-aligned with the strategies scientists have actually used to critique the aspartame research.

Our aspartame problem mirrors the problem King and Kitchener posed to students. In King and Kitchener’s version, participants were informed that there may be a relationship between additives and health outcomes, and that some research showed that
additives are unhealthy while other research showed that additives make food safer to eat. On the basis of this and other problems, King and Kitchener concluded that most college students are not evaluativists (reflective thinkers). However, students were not provided with enough details to actually reason about the problem. King, Kitchener, and colleagues have provided minimal details on students’ actual reasoning.

By focusing respondents on a specific food additive and specific methodological features, we found that students’ reasoning strategies were similar to the strategies scientists have used. We think this is indicative of sophisticated EC. Our argument (and evidence), is not new. For example, research on the Wason Selection task, a logic puzzle activity, has found that providing reasoners with contextual information significantly improves performance (Griggs & Cox, 1982). It is of no surprise that learners struggle on context lean and free, interview and likert-scale tasks, such as those used by educational psychologists and science educators. Overall, students’ strategies seemed to be well-aligned with experts’.

**Dinosaur Metabolism.** There is no consensus view on the extent to which dinosaurs are endothermic or ectothermic. Explanations are underdetermined by the data due to a significant reliance on fossil data. This issue is complicated by the fact that there is not a clear distinction between endothermic and ectothermic animals. For example, most fish are categorized as ectotherms, however larger fish, such as tuna and billfish are at least partially endothermic. Students were presented with two pieces of indirect evidence bearing on the dinosaur metabolism problem. We were interested in whether any students would note that the theories were underdetermined by the evidence.
Students who made a strong attempt at explaining the evidence did so by either suggesting that scientists were actually studying two different kinds of dinosaurs (subtyping subjects, 7 of 22 students, 32%) or were able to come up with some entirely new complex explanations (4 of 22, 18%). A significant number of students also simply said they could not explain the conflicting evidence (6 of 22, 27%).

We have no reason to believe the students in this study were aware that the distinction between endothermia and ectothermia is not straightforward; no student explicitly noted this. However, the subtyping subjects response generated by students is well-aligned with the possibility that dinosaurs may have differed in the extent to which they were endothermic or ectothermic.

A number of students generated other clever evidence-based explanations. For example, one student argued that dinosaurs were actually cold blooded (suggested by the pineal gland evidence), and that the fossils found in cold regions (suggested by the second study) could be accounted for because the region was too cold for cold-blooded dinosaurs to live there. Dinosaur fossils were found in cold regions because the cold caused them to die there. Another student suggested that cold-blooded dinosaurs in the north had a symbiotic relationship with warm-blooded animals that would keep them warm. While these responses are not normative, we do think they represent a reasonable attempt to reconcile the evidence, given students’ limited content knowledge and the limited amount of evidence. They types of responses could provide rich fodder for classroom conversation about working with multiple scientific explanations.

As noted above, when students were asked if scientists would ever be certain, 79% noted that it may be difficult to ever be certain because of a lack of evidence. Abd-
El-Khalick and colleagues have utilized a problem similar to our four problems to evaluate the extent to which students’ understand that science is theory-laden (Abd-El-Khalick’s problem focused on dinosaur extinction). Like our students, Abd-El-Khalick found that many students noted that scientists have conflicting views because of a lack of evidence. However, they concluded that students were not demonstrating an adequate understanding that scientists' disciplinary training and educational backgrounds, personal experiences, preferences, and opinions, and basic guiding assumptions and philosophies influence their perception and interpretation of the available data.”

We found that across the full four problems many middle school students noted components of scientific thinking that suggests that they do indeed have an understanding that science is theory-laden. While problems related to dinosaurs are certainly theory-laden, our evidence suggests that this nature of science issue is not particularly salient to middle school students—rather indeterminacy is due to lack of evidence is more salient.

**Pellagra.** Pellagra is caused by a niacin deficiency. We presented students with two studies that described natural interventions. In study 1, researchers found that pellagra decreased after a sewage system was installed. The researchers concluded that pellagra was caused by bacteria that spread through unclean living conditions. In study 2, in another city, researchers found that, (a) pellagra did not decrease when sanitation improved, and (b) that pellagra increased in years when the city’s economy was bad. These researchers inferred that citizens could not afford healthy food and that diet was the cause. Both pieces of evidence weakly supported their conclusions. The bacteria explanation was further cast into doubt by Study 2. The poor diet explanation relies on a rather feeble link with the evidence.
Despite the fact that bacteria explanation was weakly supported by one study and conflicted with the second study, none of the students explicitly dismissed the bacteria explanation.

The majority of students generated complex explanations that took into account the conclusions of both studies (15 of 23, 65%). None of the students explicitly dismissed the bacteria explanation.

Many of the students’ complex explanations were quite clever. For example, one student argued that bacteria were carried by particular foods:

Let’s say somehow the cows were contaminated with pellagra viruses that didn’t affect them, and then people were eating the cow meat. And the virus or bacteria that was in the meat gave them pellagra. I mean a diet does make sense, but since there’s still poor diets today and I’ve never heard of it, it doesn’t make as much sense as a virus or bacteria. But a virus or bacteria that’s carried by a diet I think could be likely (Student 27).

Another student argued that poor diet influences the body’s ability to fight off disease:

They can put these two together and then research a little more and see if poor diet, and virus and bacteria can relate to each other in some way. Because when people have a poor diet, they usually don’t have money so they’re either on the street or in bad living conditions so that can create bacteria and it can be dirty. And then usually when you have a poor diet, you get viruses and they usually don’t go away and you can’t fight it off or anything (Student 25).
The primary problem with these explanations is that students did not seem to explicitly attend to the details in the second study that weakened the bacteria/viruses explanation. The responses to this question illustrated that students can think in very complex ways, but they need more practice tracking the specific details of studies and explaining how the details of the studies connect with the alternative theories.

Students’ failure to reject the germ theory explanation could be explained by students’ familiarity with the ideas about germ theory and the relationship between diet and health (Solomon & Cassimatis, 1999). Students’ responses were likely influenced by their existing theories (Chinn & Brewer, 1993).

Although they did not attend carefully to details of the studies, a few students were sensitive to the conclusions drawn from the evidence. Seventeen percent of students attacked the conclusions drawn by the studies. For example, one student correctly argued that the relationship between building sewers and the decrease in pellagra could be non-causal. Unfortunately, this student did not mention the evidence in the second study that would have confirmed this suspicion. Another student noted that the diet conclusion was not actually based on any direct evidence. However, that student failed to recognize that the evidence in the second study weakened the claims of the first, therefore bringing the conclusion of the second study into play:

They could have looked at different aspects of what happened. The first study, they looked at how clean things were and then the second one, they looked at how clean things were but then they said that it was probably because they couldn’t get good food because they lost their jobs and that was why. They didn’t really know for a fact that the people couldn’t get good food. In study two they just assumed
that it came from a poor diet because most people got it probably couldn’t get good food (Student 23).

Overall, based on students’ responses to the pellagra problem, students certainly engaged in careful, evidence-based, evaluative thinking. Their strategies and patterns of reasoning were consistent with sophisticated epistemic beliefs. However, we thought students’ responses to the pellagra problem suggest that students need more practice attending and coordinating details of scientific studies.

Across the four problems, our interpretation of our data is that students clearly demonstrated evaluativist thinking when presented with scientific thinking. It seems unlikely that students would engage such a range of reasoning strategies if they did not have sophisticated beliefs about nature of knowing and knowledge.

Students did demonstrate some weakness in the overall quality of their reasoning. In a few cases, particularly when responding to the Frog and Pellegra problem, students neglected details of the studies that may have allowed them generate more normative strategies. Students could benefit from more practice, and perhaps more directed instruction, aimed and focusing their attention on the details of evidence and applying the best possible strategy.

**Discussion**

We have described learners’ strategies for coordinating conflicting scientific evidence. A primary goal of our research is to evaluate four predictions about learners’ reasoning and epistemic cognition. We now consider the extent to which our data supports these four predictions.
Prediction 1. Students will have few, if any strategies for coordinating scientific evidence. Prediction 1 was based on the conspicuous lack of description of participants’ reasoning in existing EC studies. EC researchers have not typically described students’ reasoning strategies and have explicitly argued that students’ do not have an understanding of the role of reasoning in sophisticated EC (e.g. King & Kitchener, 2004; Kuhn et al., 2000). Reasoning researchers have not provided detailed descriptions of students’ abilities to coordinate evidence (Chinn & Malhotra, 2002). Thus, there was no a priori basis for predicting that middle school students’ will be able to demonstrate the use of strategies for coordinating evidence.

Prediction 1 was not supported. Overall, we identified ten strategies in students’ attempts to reconcile the conflicting evidence shared with them. The four most common strategies applied by the students included (a) developing a complex model that accounted for both pieces of evidence, (b) identifying methodological flaws, (c) highlighting differences in the subjects of the studies (subtyping subjects), and (d) highlighting design differences between the studies. Students also demonstrated knowledge of the strategies that scientists could use to reconcile conflicting theories. These included (a) designing a critical study, (b) collecting an additional quantity of evidence, and (c) supporting communication between scientists. Overall, students demonstrated understanding of a wide range of reasoning strategies for coordinating and reconciling conflicting evidence.

Did students demonstrate good reasoning? It’s unclear. One limitation of this study is that we did not compare students to expert reasoners. Scientists would likely draw on additional stores of content knowledge to provide more sophisticated and
adaptive patterns of reasoning. However, it is not obvious to us that non-science, expert reasoners, such as philosophers, historians, social scientists, or lawyers will exhibit patterns of reasoning significantly different from middle school students. Comparative research will likely be fruitful and support the development of a model of reasoning that can inform instructional design.

Prediction 2. Students’ patterns and targets of reasoning will not be consistent with sophisticated epistemic beliefs. Our second prediction was based on the extant EC research and the now unsupported Prediction 1. The converging conclusions of earlier EC research are clear: The results of relatively simple interview tasks and questionnaires suggest that students do not have sophisticated epistemic beliefs. Students do not demonstrate sophisticated and adaptive patterns of reasoning across problems, while sophisticated epistemic cognizers do. Following from Prediction 1, we did not think that students would demonstrate an understanding of appropriate reasoning strategies to support sophisticated EC.

Prediction 2 was not supported. We found patterns of reasoning concordant with epistemologically sophisticated beliefs about complexity, certainty, source, and justification of knowledge. We briefly consider the concordance of students’ patterns of reasoning with each EC dimension.

Epistemologically sophisticated learners believe that knowledge is justified by evidence and through rules of inquiry. Learners with an informed understanding of the nature of science recognize that scientific claims change as new evidence is brought to bear on existing theories. The data we described showed that most students’ reasoning was targeted on the evidence provided to them. Seventy-eight percent of students
proposed a specific study (evidence) to reconcile the scientific conflict. Twenty-eight percent of students proposed that significantly more evidence should be collected by scientists to reconcile the studies. It is clear, that students had a keen understanding that scientists are reactive to evidence. Finally, all students utilized evidence-focused strategies aimed at preserving the conclusions of the studies they viewed. If students were unsophisticated, they would have blithely ignored the evidence.

Epistemologically sophisticated learners believe that the source of knowledge is reasoned inquiry and evidence. Unsophisticated learners believe that knowledge resides outside the self and is transmitted (cf. Mason et al., 2010). As we just noted in our discussion of students’ beliefs about justification, all students’ explanations were squarely focused on evidence and they were certainly engaged in reasoning about that evidence. Further, seventeen percent noted that scientists should communicate more in order to come to reconcile the conflicting evidence—suggesting an explicit awareness of the social foundations of EC (cf. Goldman, 1999; Longino, 2002). In responding to whether scientists would agree with students own explanations, 52% of students provided responses focused on the cognitive dispositions of scientific thinkers (students or scientists)—this type of response seems inconsistent with the view that knowledge is simply transmitted from one actor to another.

Epistemologically sophisticated learners believe that the certainty and complexity of knowledge are context-specific. We had a relatively direct measure of students’ beliefs about the certainty of scientific knowledge—we asked students if scientists would ever be certain for each of the problems they encountered. If students’ responses varied across problems, it would suggest that students were sensitive to contextual features of science
problems. We found that 67% of students evaluated the certainty of topic knowledge differently across problems. An additional, 13% of students stated that they were uncertain if scientists would be certain across all topics they responded to. Our data suggests that the majority of students made contextual decisions when evaluating certainty.

We think students’ views about whether scientists would agree with their attempts to reconcile the evidence is a good indication of the students’ beliefs about the complexity of knowledge. Students who believe that knowledge is complex may be skeptical about their own ability to generate a good scientific explanation. Fourteen percent of students made this exact point about the possible limitation of their knowledge without being specifically prompted. Seventy-two percent provided a combination of different types of responses or only provided uncertain or indecisive responses. Further, as we have noted throughout, learners demonstrated variability in reasoning across problems. This is also an indicator of students’ beliefs about the contextual aspects of complexity.

Overall, across the four common dimensions of epistemic beliefs, we found evidence that existing measures likely have not fully capture students’ actual EC. Students’ reasoning did not suggest that they were realists, absolutists, or relativists—based on their patterns of reasoning, our young participants could only fairly be called evaluativists or reflective thinkers. We believe that researchers should aim to develop models of EC that account for students’ patterns of reasoning.

Prediction 3. Providing learners with contextual information, such as evidence, will have no impact on their reasoning and epistemic cognition. Our third prediction was
based on research that implicitly suggests that the results of context-lean assessments will generalize to and predict performance on contextually robust and practical problems. As we have belabored above, the vast majority of research has found that students’ EC is unsophisticated.

Prediction 3 was not supported. In addressing Predictions 1 and 2, we have argued that students’ patterns of reasoning were consistent with sophisticated epistemic beliefs. Thus, our evidence-focused task did have a significant impact on the extent to which students revealed their epistemic beliefs.

At the outset of this article, we discussed a number of reasons why providing students with evidence will support their engagement in sophisticated EC. It may be instructionally important to identify which reasons best explain why our task elicited better reasoning in students. For example, if improvement in students’ responses is due to knowledge about different evidential forms (e.g. experiments, observational studies) or of specific topical knowledge, this information can be used to help curriculum designers’ make decisions about the balance between presenting general evidence forms and showing how they can be applied, or providing students with more detailed topic-specific knowledge. If students’ reasoning was improved because our problems were more motivating, this may provide important information on how to promote student engagement in the science classroom. Overall, extensions of this study may provide important insights into how to align EC research with both learning theories, such as cognitive load and motivation theories, and instructional practices.
Prediction 4. Students will demonstrate domain-level epistemic beliefs. Our final prediction was based on the extant research suggesting that students have domain-level epistemic beliefs.

Prediction 4 was not supported. Our data was consistent with emerging research suggesting that topic-level epistemic beliefs may be a better grain-size for analysis. As we discussed in addressing Predictions 1 and 2, students adaptively engaged different reasoning strategies across the four different science problems. Further, students’ judgments about the certainty of knowledge varied across problems—evidenced by the variability in the responses to whether scientists would ever be certain about each of the problems.

While the results of this study suggest that students have topic-level beliefs that may be heavily influenced by conceptual cues, it may still be useful to extrapolate from reasoners’ topic-specific patterns of reasoning to more general characterizations. For example, expert-novice comparisons may reveal more differentiated patterns of reasoning across all domains.

Instructional Implications. Based on the results of this interview study, we offer three recommendations for instruction that should be empirically explored.

First, we recommend that teachers engage in instructional activities that promote the diffusion of scientific reasoning strategies in students. Although the students surveyed in this study collectively demonstrated awareness of a wide range of strategies, it is unlikely that all individuals’ repertoires included the full corpus of strategies. Research by Anderson and colleagues (2001) suggests that student-centered critical discussion promotes the uptake of reasoning strategies by students more efficiently than teacher-
centered instruction. We have advocated explicit knowledge sharing activities in which students’ individual brainstorm, propose, vet, and adopt ideas (Pluta et al., submitted).

Second, we believe that the results of this study suggest that instruction be centered on helping students attend to important details of scientific studies. Students demonstrated a general understanding of the studies. However, they seemed to overlook more fine-grained details that may have allowed them to reconcile both studies, rather than simply dismiss one. In particular, we suggest that students are introduced to a comparison strategy in which students look for similar details across multiple studies. Often commonalities can suggest more formal relationships between studies (cf. Day & Goldstone, 2012).

Finally, we recommend that teachers and instructional designers use conflicting evidence across a wide variety of domains, both within and outside of science. This is likely to familiarize students with different disciplinary norms for dealing with underdeterminacy, error, and methodology.

Each of the recommendations is tentative, and should be examined empirically.

Summary

Learners demonstrated knowledge and application of a wide range of strategies for coordinating conflicting evidence. Their strategy use varied across scientific problems. Overall, we believe that learners’ patterns and targets of reasoning suggest that their epistemic beliefs and understanding of the nature of science are far more sophisticated than previous studies and models suggest. Further, students’ EC seems to be context-specific. We believe that studies by epistemic cognition researchers should aim to discover the basis of this.
References


King, P.M. & Kitchener, K.S. (2002). The reflective judgment model: Twenty years of research on epistemic cognition. In B. K. Hofer and P. R. Pintrich (Eds.), Personal epistemology: The psychology of beliefs about knowledge and knowing, (pp. 37-61). Mahway, NJ: Lawrence Erlbaum, Publisher.


*Developmental Psychology, 35*(1), 113.


**Themes Revisited**

The three studies described in this dissertation make a humble contribution to the science education literature. By identifying strengths and weakness in students’ reasoning and understanding of the nature of scientific inquiry, the studies serve as a guide for developing targeted instruction that has the potential to move students’ reasoning forward and support the development of sophisticated knowledge about the nature of science and scientific practices.

To review, the results of Study 1 indicated that middle school students could generate lists of epistemic criteria for evaluating scientific models that had similarities to the criteria identified by science education researchers and philosophers of science. The most commonly listed criteria referred to the clarity, pictorial form, and the explanatory function of models. Almost a quarter of the students included criteria relating to model fit with evidence. Students’ criteria provided insights into their understanding of the explanatory and descriptive goals of models; the constitutive, communicative, and epistemic features of models; and the role of evidence in supporting models. Collectively, students demonstrated familiarity with a wide range of modeling ideas that can be leveraged in instruction to promote deeper understandings of modeling practice and nature of science.

Building on Study 1, Study 2 showed that classes of students could engage in sustained discussion around proposing, vetting, and adopting epistemic criteria to be used to support model-based science instruction. In this study, we examined an instructional sequence in which students (1) generated individual lists of epistemic criteria for evaluating scientific models, (2) engaged in discussions in which they collaboratively
vetted their criteria, and (3) adopted class lists of criteria. Our analysis indicated that class adopted lists were more sophisticated than the average individual’s list, and four core criteria associated with learners’ understanding of content, epistemic development, and inquiry knowledge were present in the class lists. Learners’ initial shared understanding of the criteria (before discussion) was an important factor for predicting the content of the class lists. We argued that class lists of criteria are sophisticated enough to support student-centered instruction and promote evaluative discourse norms.

Finally, in Study 3, students interpreted the results of conflicting scientific evidence; overall, students demonstrated facility with an impressive array of reasoning strategies that have many similarities to the strategies used by scientists, as well as an emerging understanding of criteria for good evidence. We identified ten unique reasoning strategies learners used to coordinate the results of studies. The four most common strategies were (1) identifying methodological flaws, (2) identifying design differences, (3) subtyping the subjects of the studies, and (4) proposing more inclusive explanations. Learners applied these and other strategies adaptively across four different problems. Overall, learners’ patterns of reasoning were consistent with having sophisticated epistemic beliefs about the nature of science. Students focus on identifying methodological flaws and design differences indicated that students have the epistemic resources for more in-depth instructional engagement with epistemic criteria for good evidence. In addition, many of students’ reasoning strategies are well aligned with normative criteria for good models. We argued that learners’ epistemic cognition is more sophisticated than previous research suggests.

I now consider three cross-cutting themes that emerged from the studies.
**Instructional lessons learned.**

Each of the three studies provides additional information about productive starting points for instruction and areas in which students would benefit from additional targeted.

Study 1 and study 2 indicate that even a brief introduction to model activity can efficiently orient students to normative epistemic criteria for evaluating scientific models. Based on the information provided by these studies, the introduction to model task used in Study 1 and 2 could be rejiggered to promote identification and discussion of other criteria not currently supported by the task used in the present studies. Similarly, there could be greater instructional focus on supporting students’ reasoning about and with epistemic criteria.

In study 2, most of the normative epistemic criteria proposed by individuals were adopted by peers during class discussions—however, students’ reasoning about criteria could be significantly improved. This indicates the need for additional focus on promoting student elaboration by the instructor or instructional designers. One way to do this would be simply to prompt students for additional elaboration or reasons. Another way would be to give students additional opportunities to generate reasons for why their criteria are good criteria in preparation for class discussion—such as by asking students to write down their reasons before class discussions or simply think of reasons.

Overall, Study 1 and Study 2 indicate that the instructional sequence described is a promising method for promoting the adoption of evaluative discourse norms in the science classroom.

The data presented in Study 3 indicates that inquiry might support the development and elicitation of concepts indicative of a sophisticated understanding of the
nature of science. The primary benefit of direct instruction may be that it primarily introduces students with normative language to describe what they understand and believe. For example, the students who participated in Study 3 demonstrated patterns of reasoning that were consistent with holding beliefs that the complexity and certainty of science varies across different science domains. In order to perform well on assessments measuring nature of science knowledge or epistemic cognition—they will need access to new vocabulary to describe their beliefs— for example, learning to say science is “complex,” or “uncertain,” or “tentative.”

Similarly, the results of Study 3 indicate that students have significant epistemic resources that for engaging with criteria for good scientific evidence. The instructional sequence described in Study 1 and 2 could similarly be used to promote a shared classroom understanding of good scientific evidence. This type activity could also serve as a catalyst for the development of evaluative discourse norms within classrooms.

Finally, the evidence reported here indicates that high quality reasoning is impinged by students’ ability to carefully attend to the details of scientific studies that would allow them to apply better reasoning strategies. This was most visible in the results of Study 3. Future research should explore scaffolds to support students’ ability to makes sense of critical scientific details.

**Indirect Indicators of Epistemic Cognition.**

Throughout the descriptions and evaluations of the three studies, I (like many others) have expressed skepticisms of the usefulness of measures of epistemic cognition and beliefs about the nature of science based on questionnaires and interviews. The most useful classroom assessment of sophisticated epistemic cognition for teachers may be
their ability to parse student discourse around inquiry tasks for verbal (and sometimes nonverbal) clues that provide insights into students’ epistemic cognition.

A possible contribution of the three studies described here is that it might provide a set of authentic tasks that can be systematically applied within the classroom to provide additional clues related to the development of sophisticated epistemic cognition. Middle school teachers can adopt practices in which they regularly solicit students’ ideas about epistemic criteria individually (in written form) and during discussion. Similarly teachers can regularly solicit students’ strategies for evaluating multiple pieces of scientific evidence. These tasks can be integrated into teachers’ instructional quality improvement efforts around epistemic cognition in order to support on-the-fly evaluation of epistemic discourse.

**Deficit perspectives of students’ epistemic cognition and understanding of NOS**

All three studies provide information about the challenges students face when engaging in and describing epistemic practices. In many influential studies in which students’ epistemic cognition has been measured utilizing questionnaires or interviews with multiple participants, the participants have been typically provided with few specific details that can help orient them. For example, students have been asked generally about scientific models without being provided information about how the researcher or teachers are conceptualizing scientific models; in studies in which students are asked to interpret the results of scientific studies, they are not provided with details about the methodologies. Given that students have experience with a wide range of models and evidence that share features of both normative and nonnormative models and evidence—it is not surprising that students may provide less sophisticated responses. A good
example might be the ubiquity and emphasis on static models of cells or the solar system in elementary education that emphasize description, rather than using these types of models to generate explanations or predications (e.g. predications and explanations related to how do organelles interact). In this dissertation, I have argued that these details are critical for orienting students to the task.

**Coda**

This work makes a small contribution to refocusing research on epistemic cognition and reasoning in science education. It points to a need to for future working aiming to more carefully bridge different approaches to epistemic cognition research. It also points to epistemic criteria as a potentially important instructional scaffold for moving students’ thinking forward.