Changes in Students' Use of Epistemic Criteria in Model Evaluation

Na'ama Y. Av-Shalom, Hebbah El-Moslimany, Ravit Golan Duncan, and Clark A. Chinn
naama.avshalom@gse.rutgers.edu, hebbah.el-moslimany@gse.rutgers.edu, ravit.duncan@gse.rutgers.edu, clark.chinn@gse.rutgers.edu
Rutgers University

Abstract: Scientists collaboratively develop and use epistemic criteria to evaluate the products of their inquiry practices (e.g., models and arguments). Although students are able to engage in many aspects of scientific practices effectively, we know less about students' ability to develop and use epistemic criteria. In this paper we discuss a model-based inquiry intervention with seventh grade students in which students developed and used epistemic criteria when evaluating models. We explored students' use of epistemic criteria before and after the intervention. We found that students were able to generate and use a variety of epistemic criteria even before the intervention. Yet they improved in their ability to invoke fit-with-evidence as an epistemic criterion, a central criterion in modeling and argumentation. Our data suggest that students developed in their understanding of which criteria are prioritized in scientific practice—i.e., a meta-epistemic understanding of the utility and importance of epistemic criteria.

Introduction

Scientific inquiry involves understanding and using epistemic criteria to generate and evaluate epistemic products such as models and arguments (Laudan et al., 1986; Longino, 2002; Newton-Smith, 1981). Examples of epistemic criteria for good models include: models are good when they fit with evidence, have an appropriate level of detail, and address the question. The use of epistemic criteria is pervasive in scientific communities and should be included in the learning of science. Educational standards in the United States and elsewhere now advocate for engaging with scientific practices, including an understanding of the underlying epistemic commitments they reflect such as criteria for evaluating epistemic products (NGSS Lead States, 2013; Province of British Columbia, 2016). Through collaborative development and revision of epistemic criteria, students can become enculturated to the epistemologies underlying scientific practices (Berland et al., 2015).

Middle school students are capable of generating epistemic criteria for model quality even before formal instruction on modeling (Pluta, Chinn, & Duncan, 2011). Pluta et al. (2011) found that when students were tasked with comparing pairs of models, deciding which model is better and why, and then generating a list of criteria for good models, student-generated criteria spanned a wide range referring to a variety of model purposes and features, many of which aligned with the epistemic goals of scientists. While this research showed that students can generate criteria, it did not examine whether students could subsequently use these criteria to evaluate epistemic products when engaged in actual modeling and argumentation tasks, and how those criteria change as students gain facility and understanding of them through their inquiry learning experiences.

In this paper, we investigate how middle-school students' use of criteria changes during the course of a model-based inquiry program spanning 20-22 weeks. We examine change at two levels of epistemic growth: the level of practical performance and the level of meta-epistemic understanding (see Chinn, Duncan, & Rinehart, 2018). Practical performance with epistemic criteria involves the practical application of criteria such as fit with evidence and clarity of presentation to create or evaluate models. Meta-epistemic understanding involves explicitly recognizing that one has selected a model because it meets a core epistemic criterion such as it has better fit with evidence, greater clarity of presentation, and so on. Successful reasoning requires advances at both levels (Barzilai & Chinn, 2017; Chinn et al., 2018; Sandoval, 2005).

There are a number of different learning environments designed to help students engage in scientific practices (see Andriessen, 2006) and to help elucidate students' implicit understandings of epistemic criteria (e.g., Sandoval, 2003). However, only a handful of these environments have focused on students' explicit use of epistemic criteria when creating, evaluating, or revising scientific models (e.g., Schwarz & White, 2005; Stewart, Cartier, & Passmore, 2005). Therefore, we still know relatively little about whether and how students use epistemic criteria to guide their engagement in core scientific practices such as modeling. In particular we do not know much about how students' understanding and use of epistemic criteria changes as a result of inquiry instruction or how understanding at the meta-epistemic level interacts with the practical performance level (actual use of criteria).

Arguably the most central criterion for model quality is fit with evidence—how well the model accounts for the available evidence. The metacognitive understanding of the importance of fit with evidence as a criterion is what we mean by a meta-epistemic understanding of how to evaluate models appropriately. In our prior work we found that while students articulated a variety of epistemic criteria, only about a fifth of the students identified fit with evidence as a criterion (Pluta et al., 2011). Accordingly, in this study we explored in particular how students advanced in their meta-epistemic understanding of evidentiary criteria, as evidenced by both practical application of that criteria (e.g., using more evidence) and meta-epistemic use (explicitly stating that evidentiary criteria were being used) and how this growth was related to the growth in other criteria.

Focusing on explicitly mentioned epistemic criteria, we examined written assessments in which students evaluated competing models based on evidence. The individual assessment was completed before and after engagement in a several-months-long model-based inquiry curriculum in which students collaboratively developed and used public, shared epistemic criteria for evaluating models and evidence. We thus are able to expand on Pluta et al. (2011) to understand students' use of epistemic criteria and how use changes over time. We wanted to understand:

1. Which model-quality criteria do students apply in their written arguments?
2. How does students' practical use of epistemic model-quality criteria change before and after instruction?
3. How does students' meta-epistemic use of epistemic model-quality criteria change before and after instruction?

Theoretical framework

Epistemic norms arise out of collaborative meaning making by communities within the field or culture, often through argumentation (Bang & Medin, 2010; Lund, Rosé, Suthers, & Baker, 2013). In science, *epistemic understandings* (such as understanding the criteria and the processes used to generate and test models and theories) (Chinn et al., 2018; Ryu & Sandoval, 2012) have been developed, and continue to be evaluated, through discourse within the scientific community (Longino, 2002). These understandings also influence and are influenced by scientific practices (Ryu & Sandoval, 2012), such as creating models of scientific phenomena, and constructing arguments. Given the importance of modeling and argumentation in science and science education, it is important to engage students with these practices in order to help them appreciate how science is done and how scientific knowledge is developed using evidence.

A number of collaborative learning environments have been designed to develop students' argumentation skills (see, e.g., Andriessen, 2006). Studies have shown that, with proper scaffolding, students improve in their capacity to use more evidence (Suthers, Weiner, Connelly, & Paolucci, 1995), discriminate between evidence and claims (e.g., Reiser et al., 2001; Suthers et al., 1995), develop and elaborate on arguments (e.g., Reiser et al., 2001), and discuss the features of arguments (Munneke, van Amelsvoort, & Andriessen, 2003). Some of these learning environments have also been used by educators and researchers to identify students' epistemic understandings about science by analyzing students' implicit or explicit practical use of epistemic criteria when engaging in argumentation. For example, Sandoval and Millwood (2007) found that students discussed the need to warrant scientific claims with evidence, which is an important epistemic understanding in science. Sandoval (2003) also studied the interaction of students' conceptual and epistemic understandings of science as they used an inquiry learning environment (BGuILE). Sandoval focused on students' epistemic understanding, which students may not state explicitly but which they use when engaging in scientific practices such as exploring scientific phenomena and using data. He found that students began to develop abilities in using data when theorizing about scientific phenomena. Students were especially able to generate causal mechanisms to explain data, which is a commonly used scientific practice that may help students make connections between evidence and claims (Sandoval, 2003). Overall, students are able to engage in many of the epistemic aspects of argumentation with appropriate scaffolding.

However, students continue to have difficulties with some epistemic aspects of model evaluation and argumentation, particularly with regards to sufficiently privileging fit-with-evidence as a core epistemic criterion (Munneke et al., 2003). For example, students have difficulty adapting their views about the nature of empirical results in science (Sandoval & Morrison, 2003), such as persistently assuming that empirical results are proof rather than evidence for theories, and struggle to attend to alternative interpretations of evidence. Struggling with the purpose of evidence and with attending to alternative interpretations of evidence suggests that students are not yet able to evaluate their knowledge in a sophisticated manner, and is indicative of the need for further development of their epistemic understanding for and of evidence evaluation practices. These difficulties may be due in part to the lack of emphasis on metacognitive epistemic understanding, that is, why these epistemic criteria matter. Most of these environments did not provide opportunities for students to evaluate their epistemic understandings.

Moreover, assessing students' metacognitive understanding of epistemic criteria still poses a research challenge; for example, commonly used surveys of epistemic beliefs are decontextualized and are thus not appropriate measures of students' capacity to understand and use epistemic criteria (Sandoval, 2005). Berland et al. (2015) developed the Epistemologies in Practice framework for analyzing students' understanding of, and engagement with, scientific practices, arguing that understanding and engaging are intertwined processes. They focused on analyzing student discourse to elucidate the epistemic criteria that students use, often implicitly, when constructing scientific knowledge, and on community development and usage of these products in scientific contexts. This provides a method of examining community use of epistemic criteria to guide knowledge production, but because there is no distinction between implicit and explicit use of epistemic criteria it does not afford the ability to identify students' metacognitive understanding alongside their practical use.

Through our analysis we aim to address several gaps in existing research. First, we are interested in students' meta-epistemic use of shared epistemic criteria (i.e., explicitly stating criteria) and how this relates to their use of these criteria at the practical level of using criteria to evaluate models. Given that epistemic criteria were made explicit in the intervention and were assessed (students collectively developed, discussed, and revised public, community epistemic criteria for evaluating models and evidence), we investigated whether students would explicitly reference these criteria in their arguments about models and evidence. Second, we examined the change in students' use of epistemic criteria before and after an intervention in which they had opportunities to use, discuss, and refine these criteria.

Intervention

The middle-school students in this study participated in a life-science model-based inquiry curriculum over the course of several months. The curriculum scaffolded students as they developed, evaluated, and revised models based on evidence. Students also engaged in written and verbal argumentation about the models, evidence, and criteria. The instructional modules were co-designed by the researchers and teachers. Topics included natural selection, genetics, and cell organelles. The curriculum involved individual, group, and class activities. As part of the intervention, students developed, discussed, and revised class criteria lists for what makes good models and used these lists in their creation and evaluation of models. Students were first tasked with developing these criteria after an introductory activity in which they were exposed to several different kinds of models and representations which they were asked to discuss and evaluate in pairs. Students first developed criteria individually, followed by a class discussion in which students collaboratively developed and agreed on a class list. Among their model evaluation criteria students brought up issues of evidentiary support, pertinence to the question at hand, clarity (including using graphs and images), appropriate levels of complexity, and others. These criteria were publically displayed in the classroom and students and teachers referred to them as they engaged in modeling activities. For example, when working on revising a model, students might discuss the areas in which the model might fail to meet the criteria on the list and try to adapt it. The lists were also periodically revised and refined by the class as students developed and refined their understanding of criteria throughout the intervention.

In a typical activity in the unit, the students usually considered two or more competing models explaining the same phenomenon (e.g., model for the function of the nucleus); students evaluated the models and, in some cases, also developed or revised models. Students were given three to six pieces of evidence of varying quality to use in their model evaluation. For example, in the lesson about the cell organelles, students worked with evidence about the number of mitochondria in the muscles of different birds in order to determine the function of mitochondria. They used this evidence in tandem with competing models to determine the function of mitochondria, as well as to learn to evaluate the quality of evidence and models. Evidence was usually presented to groups through computer-based animations and simulations, and also through written reports and hands-on experiments. The evidence was developed to be engaging to students (Chinn et al., 2018) and was used to help visualize complex mechanisms more clearly. In alignment with guidelines from previous research (Rinehart, Duncan, Chinn, Atkins, & DiBenedetti, 2016), evidence varied in complexity, quality, sourcing, presence of data, and relevance to the models. For example, evidence may have been collected by a reliable source using sound methods, or a less reliable source with questionable methods. We planned evidence so that it would problematize students' evidence evaluation criteria. We have found that this variation helps support richer evidence-quality discussions (Rinehart et al., 2016); in addition, variation is important because it better approximates the range of evidence that people are exposed to in their daily lives.

Activities also included different kinds of prompts including comprehension checks, questions about the quality of pieces of evidence, and questions about the relationship between evidence and the models.

Throughout this process of evaluating evidence and models, students collectively developed and revised criteria lists for model evaluation and practiced using them alone and in groups.

Two main scaffolds supported students in these tasks, in particular evaluating evidence quality and linking evidence to models. First, students used a 0-3 scale to evaluate evidence quality, with zero being very bad evidence so that the conclusions cannot be believed and the evidence should be ignored, and a three meaning that the evidence was of high quality and its conclusions can be believed. A second scaffold was the model-evidence link (MEL) diagram, a chart in which students used arrows to signify the relationship between each piece of evidence and model. Students used five relationship arrows signifying that the evidence highly supports, supports, highly contradicts, contradicts, or is neutral to the model (see Figure 1).

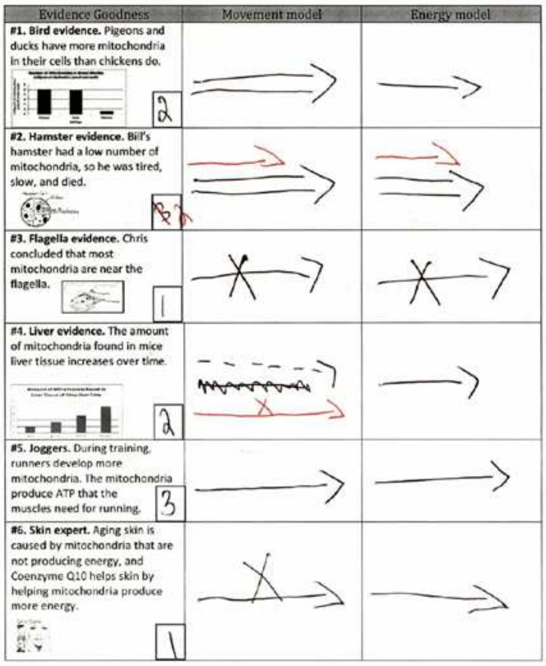

Figure 1. Model-Evidence Link Diagram.

Method

Research context and participants

The study included data from 204 seventh grade students in 15 classrooms taught by three teachers in a suburban middle school in a township in the Northeast of the United States. Based on the state report card, 31.2% of the students in the school were Asian, 5.0% Black, 7.2% Hispanic, 56.1% White, and 0.5% other. 14.1% qualified for free or reduced lunch, and the performance of this school was above the state average.

In this paper we report on our analysis of the pre and post modeling and argumentation assessment. We developed two comparable versions of the assessment: one about why we feel muscle pain 48-72 hours after exercise (MP), and the other about why leaves fall off trees in autumn (FL). These were counterbalanced between the pretest and posttest (i.e., some students received MP as a pretest and FL as posttest, whereas others completed the assessments in the opposite order). The assessment introduced students to two competing models. In FL the model better supported by the evidence—which we will call the "correct" model—was Model A. This model proposed that shorter days induced the production of a poisonous chemical that killed cells in the leaf stalk and caused the leaves to fall off, whereas the model less supported by the evidence (the "incorrect" model)—Model B—proposed that below-freezing nights caused ice crystals to form in the leaves, killing leaf cells and causing the leaves to fall (see Figure 2). In MP the incorrect model, Model A, proposed that lactic acid builds up in the cells causing them to swell and push against the nerves resulting in pain, whereas the correct model, Model B, proposed that the muscle fibers are damaged during exercise and the process of repair is painful. Students were provided with five pieces of evidence to help them decide which model is better. The first two pieces of evidence reiterated various aspects of the phenomenon for both FL and MP. The other three pieces of evidence supported, but to different extents, the correct model in both assessment versions. Students were then prompted to answer: “Which do you think is the better model for the problem? Write at least three (3) detailed reasons for your answer.”

Data analysis

Pluta et al. (2011) developed a coding scheme for model-quality criteria based on the criteria seen in students’ class lists, which we expanded to reflect additional criteria present in students’ essays (see Table 1), as well as further adapted to capture evidence-quality criteria. Model-quality criteria and evidence-quality criteria were the justifications that students gave in support or contradiction of a model or associated piece of evidence which are based on epistemic reasons (general reasons for model or evidence quality, such as “has sufficient details”) rather than empirical reasons (stating specific pieces of evidence or prior knowledge). The criteria included ones relating to empirical considerations (e.g., supported by most of the evidence, includes a sequence of steps), communicative considerations (e.g., clarity of the model, presence of diagrams or charts), pertinence (e.g., model answers a question), and others. We defined the model that students identified as being better as their chosen model. Coding was done by a pair of coders. Disagreements were settled through discussion.

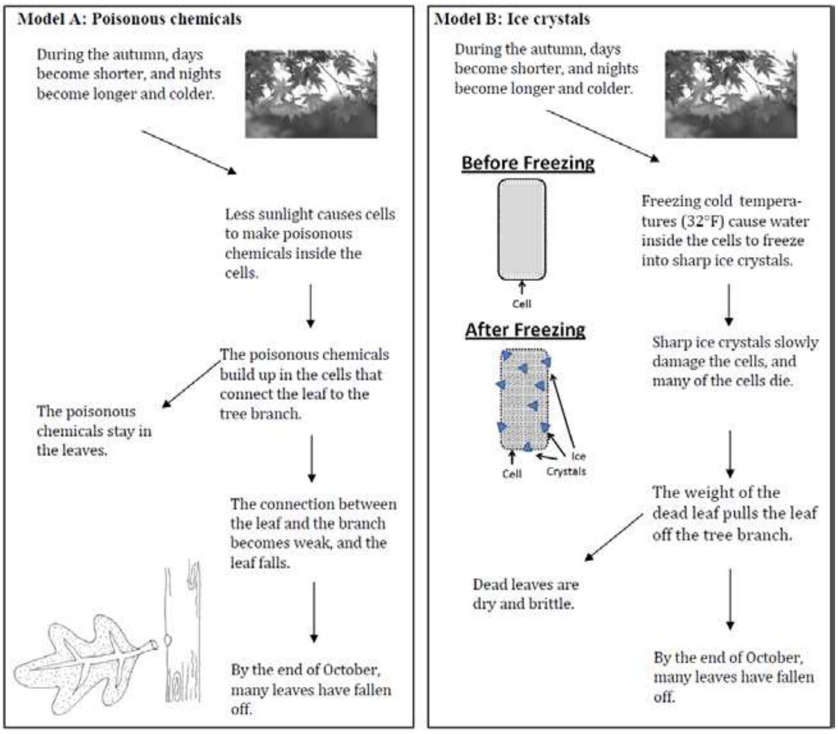

Figure 2. Falling Leaves Alternative Models.

Results and discussion

At the beginning of the model-based inquiry curriculum, students collaboratively developed class lists of criteria for good models that included a range of reasonable criteria. All of the 7th grade classes involved in this study had some form of “fit with evidence” as a criterion on their class criteria lists. Thus, the intervention produced the expected class sets of public criteria generated by students for their own use as a community.

At the practical level of using epistemic criteria on the pretest and posttest, we first note the extent to which students actually used evidence in their arguments on the pretest and posttest. We found that 49% of students included at least one piece of evidence in their arguments on the pretest, which increased to 80% on the posttest ($p<.05$). The average number of pieces of evidence used on the pretests was 0.94; on the posttests, it was 1.96 ($p<.05$) (pretest: no evidence: 51%, one piece of evidence: 20%, two: 14%, three-five: 15%; posttest: no evidence: 21%, one: 21%, two: 25%, three-five: 36%).

We also found a significant increase in students’ explicit noting of fit with evidence as a criterion for good models (indicating a meta-epistemic understanding) in their arguments (pre: 15%; post: 35%) (see Table 1), suggesting that students’ meta-epistemic understanding about the importance of evidence as a criterion was improving along with their practical performance (ability to use evidence). The fit-with-evidence criterion included five sub-categories, reflecting the extent to which the evidence set supported/contradicted the models.

Of the five sub-categories, we found a significant increase from pre (5%) to post (22%) in the number of students who specifically noted that more/most evidence supported their chosen model.

Although students also frequently mentioned several other criteria explicitly (Table 1 presents the most frequent categories), only the fit-with-evidence category showed a statistically significant increase in explicit use from pretest to posttest. This suggests that students improved in their understanding of the importance of evidentiary support as a key epistemic criterion for model goodness. Further, they were also able to attend to the proportion of supporting evidence pieces within a set (e.g., noting that most or more of the evidence supported their model), and to the importance of the absence of counterevidence for their chosen model (i.e., that none of the evidence goes against their chosen model). A model supported by more of the evidence and without any counterevidence to contradict it was viewed as a superior model.

Table 1: Categories used by over 10% of students in pre and/or post

Category & Definition	**Examples**	**Pre %**	**Post %**
Fit with Evidence The student refers to the degree to which evidence is included/supports/contradicts a model.	"I think lactic acid model is better. It supports and has stuff from the evidence." "This is clearly evident as most to all of the evidence supports this reasoning." "The other model has no evidence supporting it." "I believe the Poisonous Chemicals is the better model. None of the evidences below seem to go against the model."	15	35
Makes Sense The student refers to the degree to which the model makes sense.	"It makes sense." "I firmly believe that the lactic acid model is the better. First of all, it makes more sense than the other model." "Explanation 1 doesn't really make a lot of sense."	19	19
Explains Student refers to the extent to which a model explains/has an explanation.	"I think that the Lactic Acid Model is a lot better because it explains everything." "The Ice Crystal model shows how the leaves fall off"	14	11
Use of Visuals (picture/diagram/charts) The student refers to the quality/number of pictures/diagrams/charts in a model.	"I also think it's a better because it shows more understandable pictures." "I believe that the Lactic Acid Model is better because it has a before and after picture" "It has more diagrams than the other model"	7	10
Realistic The student refers to the degree to which a model is realistic.	"Next, the poisonous chemicals is better because it seems more realistic." "Ice Crystals may not seem realistic in some areas to lose leaves."	12	6

In terms of criteria other than fit with evidence, there was not a statistically significant difference in the percent of students who used model-quality criteria as part of the justification for choosing a particular model (pre: 57%; post: 65%). There was also no statistically significant difference in average number of model-quality criteria provided explicitly (pre: 0.59; post: 0.74). Although the students engaged with model-quality criteria throughout the intervention, there was little change on these dimensions overall (both the percent of students mentioning criteria and the average number of criteria mentioned) between pre and post. Thus, the overall picture that emerges is that students initially (at pretest) used a range of epistemic criteria at the meta-epistemic level, but most of these were not evidence-related. The model-based inquiry intervention produced a very specific effect: *it increased both practical and meta-epistemic use of fit with evidence as an epistemic criterion*.

It is important to note that there are a variety of other aspects of arguments that students learned about, such as the use and description of evidence, explaining connections between evidence and models, and providing reasons that justify the link between the model and its supporting evidence. See Table 2 for an example of good essays with and without model-quality criteria. Given that students learned about all of these different aspects, it is reasonable that many of them decided to focus on aspects other than criteria, such as explaining the conclusions of the evidence or noting whether and how evidence supports or contradicts a model. It is thus encouraging that the number of criteria mentioned remained stable and suggests that students are developing both their practical use and meta-epistemic understanding of the importance of model-quality criteria. Furthermore, students began the intervention already being able to identify epistemic criteria but using evidence and articulating fit with evidence as a criterion at fairly low rates so this selective change, rather than an overall increase in the usage of all criteria, suggests that students improved in their meta-epistemic understanding of which factors are prioritized in scientific practice.

Table 2: Good essays including and not including model-quality criteria: both show practical engagement with the material (e.g., both attended to key pieces of evidence and described their relationship with the model)

With model-quality criteria	I think the Muscle fibers model is better for three reasons. First *more evidence supports* this model. Evidence 5 and 4 support model B because they both contradict model A. Second, Model B talks about how the person gets stronger after exercise which *would make sense*; model-A however doesn't talk about that so *it isn't as realistic* as M-B. Lastly, Model B is better because it says there is damage to muscle fibers. Many times during exercise we pull muscles and damage them. *Model B explains these but Model A doesn't.*	In this essay, the student used four model-quality criteria, fit with evidence, makes sense, realistic and explains (noted in italics). This student also engaged in other aspects of argumentation, including discussing the specific relationship of evidences 5 and 4 to the model, comparing information in the two models, and giving examples from prior knowledge.
Without model-quality criteria	I think the poisonous chemicals model is best because in Evidence 5; it showed that even though a whole month without below freezing took place, leaves were still falling off, which contradicts Model B, the ice crystals model. Model B states that when it's below freezing temperatures, ice crystals form and weigh the leaf down which causes it to fall.	In this essay the student did not use any model-quality criteria. However, the student described a key piece of evidence (evidence 5) and explained its relationship to the alternative model. She then described the relevant part in the alternative model.

Conclusion and contribution

This research indicates that, with curricular scaffolding, students are able to improve in both their practical use of epistemic criteria for model quality and their meta-epistemic understanding of the role of these criteria in model evaluation and argumentation. When students began the intervention, each class developed a list of community epistemic criteria that included many criteria that are used by scientists. On the pretest, they demonstrated that they were also able to use those criteria to justify their arguments. The students continued to use epistemic criteria at similar levels in their posttests. There was a significant increase only in students' meta-epistemic articulation of fit with evidence as an epistemic criterion. There was a corresponding increase in the practical use of evidence in their argumentation. Scientists use evidence as a primary means to develop, evaluate, and justify models, and yielding to evidence is a central tenet of science knowledge building. Thus, it is crucial that students grow to understand this epistemic cornerstone. The students' selective increase in using and articulating the need for evidence suggests that, indeed, throughout the intervention students refined their meta-epistemic understanding of criteria, raising the importance of fit with evidence as a core criterion.

Students may have used their meta-epistemic understanding to regulate their practical performance, increasing use of evidence as they came to appreciate fit with evidence as a critical criterion for good models. Students started with a broad awareness of a range of epistemic criteria, but their initial criteria significantly underrated the importance of evidentiary criteria. The model-based learning curriculum produced a highly targeted improvement in both the practical use and meta-epistemic use of evidentiary criteria; given the importance of evidentiary support in scientific practice (e.g., scientists use evidence as a primary means to develop, evaluate, and justify models), the move towards the use of evidence and evidentiary criteria suggests a growth in students' meta-epistemic understanding of and ability to engage in scientific practice.

Understanding more about students' decisions about which epistemic criteria to attend to helps elucidate more about the complex processes governing how students evaluate and use scientific information. This will help teachers, researchers, and others in the education community develop learning environments which will foster the skills needed to aptly engage with science in real-world settings. It is particularly interesting to note that students may come to classrooms with the resources to contribute many of the building blocks needed to engage with scientific practices, such as developing epistemic criteria and using evidence. However, classroom interventions may help them reflect on and refine their knowledge in order to better understand the reasoning that scientists use when they engage in these processes.

References

Andriessen, J. (2006). Arguing to learn. In R.K. Sawyer (Ed.). *The Cambridge handbook of the learning sciences* (pp. 443-460). New York, NY: Cambridge University Press.

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. Science Education, 94(6), 1008–1026. https://doi.org/10.1002/sce.20392

Barzilai, S., & Chinn, C. A. (2017). On the Goals of Epistemic Education: Promoting Apt Epistemic Performance. *Journal of the Learning Sciences*, 1-37.

Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2015). Epistemologies in practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching*, *53*(7), 1082-1112.

Chinn, C. A., Duncan, R. G., & Rinehart, R. W. (2018). Epistemic design: Design to promote transferable epistemic growth in the PRACCIS project. In E. Manalo, Y. Uesaka, & C. A. Chinn (Eds.), *Promoting spontaneous use of learning and reasoning strategies: Theory, research, and practice for effective transfer* (pp. 242-259). New York: Routledge.

Laudan, L., Donovan, A., Laudan, R., Barker, P., Brown, H., Leplin, J., Thagard, P., & Wykstra, S. (1986). Scientific change: Philosophical models and historical research. *Synthese*, 69, 141–223.

Linn, M. C., Clark, D., & Slotta, J. D. (2003). WISE design for knowledge integration. *Science Education*, 87(4), 517–538. https://doi.org/10.1002/sce.10086

Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press.

Lund, K., Rosé, C. P., Suthers, D. D., & Baker, M. (2013). Epistemological encounters in multivocal settings. In *Productive multivocality in the analysis of group interactions* (pp. 659-682). Springer, Boston, MA.

Munneke, L., van Amelsvoort, M., & Andriessen, J. (2003). The role of diagrams in collaborative argumentation-based learning. *International Journal of Educational Research*, 39(1), 113-131.

Newton-Smith, W. (2002). *The rationality of science*. London: Routledge.

NGSS Lead States. (2013). Next Generation Science Standards: For States, By States.

Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching*, 48(5), 486-511.

Province of British Columbia. (2016). BC's New Curriculum.

Reiser, B. J., Tabak, I., Sandoval, W. A., Smith, B. K., Steinmuller, F., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. *Cognition and instruction: Twenty-five years of progress*, 263-305.

Rinehart, R. W., Duncan, R. G., Chinn, C. A., Atkins, T. A., & DiBenedetti, J. (2016). Critical Design Decisions for Successful Model-Based Inquiry in Science Classrooms. *International Journal of Designs for Learning*, *7*(2).

Ryu, S., & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education*, *96*(3), 488-526.

Sandoval, W. A. (2003). Conceptual and Epistemic Aspects of Students' Scientific Explanations. *Journal of the Learning Sciences*, 12(1), 5–51. https://doi.org/10.1207/S15327809JLS1201_2

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education*, 89(4), 634–656. https://doi.org/10.1002/sce.20065

Sandoval, W. A., & Millwood, K. A. (2007). What can argumentation tell us about epistemology?. In S. Erduran and M. P. Jiménez-Aleixandre, *Argumentation in science education* (pp. 71-88). Springer Netherlands.

Sandoval, W. A., & Morrison, K. (2003). High school students' ideas about theories and theory change after a biological inquiry unit. *Journal of research in science teaching*, *40*(4), 369-392.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and instruction*, *23*(2), 165-205.

Stewart, J., Cartier, J. L., & Passmore, C. M. (2005). Developing understanding through model-based inquiry. *How students learn*, 515-565.

Suthers, D., Weiner, A., Connelly, J., & Paolucci, M. (1995, August). Belvedere: Engaging students in critical discussion of science and public policy issues. In Proceedings of the 7th World Conference on Artificial Intelligence in Education (pp. 266-273).

Acknowledgments

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