Received: 12 August 2016 | Revised: 2 September 2017 | Accepted: 5 September 2017

DOI: 10.1002/sce.21315

WILEY

LEARNING

High school students' evaluations, plausibility (re) appraisals, and knowledge about topics in Earth science

Doug Lombardi | **Elliot S. Bickel** | **Janelle M. Bailey** | **Shondricka Burrell**

Department of Teaching & Learning, Temple University, Philadelphia, PA, USA

Correspondence
Doug Lombardi, Department of Teaching & Learning, 450 Ritter Hall, 1301 Cecil B. Moore Av., Philadelphia, PA 19122, USA.
Email: doug.lombardi@temple.edu

Abstract

Evaluation is an important aspect of science and is receiving increasing attention in science education. The present study investigated (1) changes to plausibility judgments and knowledge as a result of a series of instructional scaffolds, called model–evidence link activities, that facilitated evaluation of scientific and alternative models in four different Earth science topics (climate change, fracking and earthquakes, wetlands and land use, and the formation of Earth's Moon) and (2) relations between evaluation, plausibility reappraisal, and knowledge. Repeated measure multivariate analyses of variance (MANOVAs) showed that participants' plausibility judgments shifted toward scientifically accepted explanations and increased their knowledge about relevant Earth science topics after participating in the activities. Structural equation modeling revealed that 10% of the postinstructional knowledge scores were related to participants' evaluations, above and beyond background knowledge, which accounted for 26% of the variance. The activities used in this study may help students develop their critical thinking skills by facilitating evaluation of the validity of explanations based on evidence, a scientific practice that is key to understanding both scientific content and science as a process. However, results from the study were modest and suggest that additional research, from both theoretical and empirical perspectives, may be warranted.

KEYWORDS
critical thinking, Earth science, evaluation, knowledge, plausibility

1 | INTRODUCTION AND PURPOSE

Critique and evaluation are central to the scientific enterprise. Recently, *A Framework for K–12 Science Education* identified critiquing, arguing, and analyzing as evaluative processes that are foundational to science (National Research

Council [NRC], 2012). Evaluation is included in scientists' questioning and observing, collecting data and experimenting, and imagining and constructing of explanations about all aspects of phenomena. "Indeed, the only consistent characteristic of scientific knowledge across [all] disciplines is that scientific knowledge is open to revision in light of new evidence" (i.e., through evaluative processes that connect evidence to explanations; NGSS Lead States, 2013, Vol. 2, p. 96). Use of evaluation and critique to construct evidence-based explanations goes beyond the notion that scientific knowledge is tentative (Lederman, 1992) and more accurately reflects an evaluative stance about the nature of science (Allchin, 2011).

K–12 science education has underemphasized the role of critique and evaluation in constructing understanding of scientific topics. Although *A Framework for K–12 Science Education* lists evaluation in the title of one of its eight scientific practices (i.e., "obtaining, *evaluating*, and communicating information," NRC, 2012, p. 3; emphasis ours), we agree with Ford's (2015) position that all scientific practices are based on "processes of perpetual evaluation and critique that support progress in explaining nature" (p. 1043). Furthermore, the *Next Generation Science Standards* (NGSS) clearly shows explicit connections between students' engagement in evaluation and learning about the scientific practices (NGSS Lead States, 2013, Vol. 2, pp. 67–78). Henderson, MacPherson, Osborne, and Wild (2015) argue that when students engage in critique and evaluation in the science classroom they gain a deeper understanding of both *what scientists know* (i.e., scientific content) and *how scientists know what they know* (i.e., scientific epistemology). In other words, explicit and purposeful evaluation when considering explanations about phenomena may facilitate deep understanding, especially when students reflect on their personal epistemic judgments and compare how scientists actually construct and reconstruct knowledge.

One epistemic judgment that is active in both laypersons' (e.g., students, the public) and scientists' thinking is plausibility. Lombardi, Nussbaum, and Sinatra (2016) have characterized plausibility as a tentative and provisional judgment about the truthfulness of an explanation. Individuals often make judgments about plausibility implicitly and automatically without much conscious thought. But individuals can also make judgments about plausibility that are explicit and purposeful. Furthermore, scientific evaluations about an explanation's plausibility can facilitate students' shifts toward greater scientific understanding. In a recent study, middle-grades students deepened their understanding of the fundamental scientific concepts related to climate with a short duration (two class periods) activity that prompted students to simultaneously weigh the connections between lines of scientific evidence and alternative explanations of the causes of current climate change (i.e., human-induced climate change, which is the scientifically accepted explanation, and Sun-induced climate change, which is a plausible but nonscientific explanation) (Lombardi, Sinatra, & Nussbaum, 2013). Interestingly, these learning gains persisted 6 months after this short instructional activity and related to appreciable shifts in students' plausibility judgments toward the scientific explanation. Lombardi and colleagues (Lombardi et al., 2013; Lombardi, Nussbaum, & Sinatra, 2016) speculated that the study's instructional activity promoted students' critical evaluations.

The present study examined Lombardi and colleagues' (Lombardi et al., 2013; Lombardi, Nussbaum, & Sinatra, 2016) speculation about the potential relations between students' evaluations, plausibility judgments, and knowledge of scientific topics. As part of a 3-year project funded by the U.S. National Science Foundation, the present study represents a focused quantitative analysis of changes in plausibility judgments and content knowledge observed in the curriculum focused on instructional scaffolds called model–evidence link (MEL) activities (Chinn & Buckland, 2012; Lombardi et al., 2013). MELs facilitate students' evaluations of the connections between lines of evidence and alternative explanations. We specifically examined high school students' repeated use (four times over the course of a 10-month school year) of MELs developed by our project team that cover the following socioscientific topics in Earth science: (a) causes of current climate change, (b) connections between fracking and earthquakes, (c) wetlands protection and land use, and (d) formation of Earth's Moon. For the present study, our research questions were:

1. How do plausibility judgments and knowledge change when engaging in an instructional activity that facilitates high school students to think critically about controversial and/or complex Earth science topics?[1]
2. What are the relations between evaluation, plausibility reappraisal, and postinstructional knowledge, above and beyond background knowledge?

Conducting our research in Earth science provided a rich venue to study topics that are important and complex phenomena (e.g., climate change, fracking, wetlands protection, and formation of Earth's Moon). Studying these topics may be especially relevant for investigating evaluation of alternative explanations, where there may be a gap between the explanations that laypersons (students, in this particular study) and scientists find plausible (i.e., *a plausibility gap*; Lombardi et al., 2013). These topics also include fundamental principles from areas covered in many high school Earth science courses, including geophysics, water resources, climate and weather, and astronomy. Prior to detailing our study's methods and results, we first discuss how we grounded our research in literature about scientific evaluation, plausibility appraisals, and science content knowledge.

2 | THEORETICAL FRAMEWORK

Central to our theoretical framework is the characterization of scientific thinking as the "consciously controlled evaluation of [explanations] in the light of [evidence]" (Kuhn & Pearsall, 2000, p. 126). By explanations, we are specifically referring to accounts of how phenomena unfold that may lead "to a feeling of understanding in the reader/hearer" (Brewer, Chinn, & Samarapungavan, 1998, p. 120; see also Braaten & Windschitl, 2011). Scientific explanations may include fully developed theories and models, as well as facets of theories and models containing essential kernels of theory (Giere, 2010; Salmon, 1994). *A Framework for K–12 Science Education* says that "Scientific explanations are explicit applications of theory to a specific situation or phenomenon, perhaps with the intermediary of a theory-based model for the system under study" (NRC, 2012, p. 52). One important criterion for the validity of a scientific explanation is its plausibility, especially the plausibility of the explanation relative to alternatives (Hogan & Maglienti, 2001). Students and the general public also engage in evaluations based on the plausibility of explanations, but often do so implicitly (i.e., without much or any thought; Lombardi, Nussbaum, & Sinatra, 2016). The present study builds upon these ideas. Specifically, the present study uses a recent theoretical model that views evaluation as a central component in the dynamic process of students' explicit evaluations and reappraisal of an explanation's plausibility (Lombardi, Nussbaum, & Sinatra, 2016). This model also posits that evaluations and plausibility reappraisal may facilitate deeper knowledge about science. As a foundation for the present study, the following subsections provide more details on the theoretical connections of evaluation, plausibility judgments, and knowledge.

2.1 | Evaluation

Some think of evaluation as a post hoc activity. For example, a teacher may ask students to evaluate whether their classroom experiment was valid and whether reliable conclusions emerged from data collected. However, we view evaluation as an ongoing process central to virtually all scientific activities. Evaluations generally involve comparison (e.g., between criteria of validity and selection of reliable evidence; between alternative explanations about a phenomenon; Duschl & Osborne, 2002; Ford, 2015). In the process of constructing valid scientific knowledge, evaluation should also involve critique. Combining critique with evaluative processes results in what some have called *critical evaluation* (Duschl & Osborne, 2002), an essential mode of critical thinking (Bailin, 2002). Helping students to become more critically evaluative as they learn about science will potentially equip them to be more scientifically literate. In fact, the NGSS states that engagement in and knowledge of scientific and engineering practices, as well as disciplinary core ideas, "should enable students to evaluate and select reliable sources of information and allow them to continue their development well beyond their K–12 school years as science learners, users of scientific knowledge, and perhaps also as producers of such knowledge" (NGSS Lead States, 2013, Vol. 1, p. xv).

The NGSS includes evaluation in several performance expectations (i.e., what students should be able to do). With frequent mention of evaluation in the NGSS, there is a need for greater understanding about "how instruction should support, over time, students' abilities to participate in it" (Ford, 2015, p. 1047). One way to promote evaluative processes may be to explicitly engage students in judgments about knowledge and knowing (i.e., epistemic judgments, such as plausibility).

2.2 | Appraisal and reappraisal of plausibility

Plausibility is a tentative epistemic judgment conducive to knowledge construction and reconstruction both in science and in science classrooms. For example, researchers have implicated plausibility judgments in facilitating co-construction of knowledge in discourse associated with collaborative argumentation (Driver, Asoko, Leach, Scott, & Mortimer, 1994; Duschl, 2007; Nussbaum, 2011). Researchers have also proposed that plausibility may be an important judgment involved in the conceptual change process (Dole & Sinatra, 1998; Pintrich, Marx, & Boyle, 1993; Posner, Strike, Hewson, & Gertzog, 1982).

Until recently, little work has attempted to empirically investigate the role of plausibility judgments in knowledge construction and reconstruction in science learning and teaching. This gap in understanding prompted an initial study by Lombardi and Sinatra (2012), which revealed a substantial connection between undergraduate students' plausibility judgments about human-induced climate change and knowledge of weather and climate distinctions. A follow-up study revealed that background knowledge, topic emotions (i.e., anger and hopelessness), and epistemic motives (i.e., decisiveness) predicted plausibility judgments about human-induced climate change in preservice and inservice teachers (Lombardi & Sinatra, 2013). These two studies motivated Lombardi et al. (2013) to investigate the connections between cognitive evaluations about scientific explanations and appraisals of plausibility about these explanations. In this study, middle school students first learned about the plausibility judgment and its tentative nature (i.e., plausibility is a judgment that scientists make about the potential truthfulness of one explanation compared to another; the judgment may be uncertain; and scientists do not have to be committed to the decision and may change their plausibility judgments about explanations). The students then engaged in a task explicitly asking the students to rank (from most to least important) the different ways in which lines of evidence and scientific explanation may be connected (i.e., the evidence strongly supports the explanation, the evidence supports the explanation, the evidence contradicts the explanation, and the evidence has nothing to do with the explanation), and how this connection influences their plausibility judgments. Finally, students participated in an activity where they explicitly considered the plausibility of two explanations about a phenomenon (i.e., a scientific accepted explanation and a compelling alternative explanation). Students specifically weighed the strength of connections between lines of evidence and alternative explanations. Through this series of activities, Lombardi et al. (2013) argued that students come to understand science as a way of knowing because they "come to appreciate that alternative interpretations of scientific evidence can occur, that such interpretations must be carefully scrutinized, and that the plausibility of the supporting evidence must be considered" (NRC, 2012, p. 251). Furthermore, by considering the plausibility of alternative explanations, students constructed and reconstructed knowledge in a way that scientists do, which in turn informed Lombardi, Nussbaum, and Sinatra's (2016) model of plausibility judgments in conceptual change (PJCC).

The recently-created PJCC incorporates philosophical and psychological perspectives. As such, the PJCC is a detailed explanatory description of the factors that potentially form judgments of plausibility. The nexus of the PJCC involves a theoretical account of how explicit and critical evaluations about novel explanations and background knowledge may influence reappraisals of plausibility, which in turn might influence knowledge reconstruction. In their development of the PJCC, Lombardi, Nussbaum, & Sinatra (2016) also discussed the meaning of plausibility from both the philosophy of science and science studies perspectives. Specifically, Lombardi, Nussbaum, & Sinatra (2016) discussed how scientists incorporate plausibility judgments into their evaluations of explanations (see, e.g., Brewer et al., 1998). In short, plausibility as an epistemic judgment of potential truthfulness is an outcome of individuals' evaluations.

Lombardi, Danielson, and Young (2016) recently tested this theoretical relation between evaluation and plausibility judgments. In this study, undergraduate students who read a refutation text (i.e., a text that identified and directly refuted common misconceptions) activated their abilities to be critically evaluative. Critical evaluation, in turn, shifted students' plausibility judgments and reconstructed their knowledge toward the scientific explanation that human activities are causing current climate change. In comparison, reading an expository text (i.e., a text that simply described and informed without refutation, a mode commonly used in textbooks) did not activate critical evaluation. In the expository text group, students retained existing, nonscientific conceptions. However, because refutation texts are not used consistently in science education (Sinatra & Broughton, 2011), we have also wondered about more ecologically valid

instructional activities that could promote critical evaluation, plausibility reappraisal, and greater knowledge about scientific topics—the focus of the present study.

2.3 | Instruction to promote evaluation and plausibility reappraisal

We acknowledge here that our use of the term "evaluation" in the context of science teaching and learning may be a bit confusing. Many researchers and educators encourage students to formally and informally evaluate understanding about *what* is being learned, as well as *how* it is being learned (i.e., as suggested by the 5E instructional model; Bybee, 2009). In this regard, researchers, educators, and students view evaluation as a key component of self-regulated learning and metacognition (Zimmerman, 1990). However, we want to stress that this study focuses on evaluation as a process that is foundational to scientific practices (i.e., critical evaluations are at the core of scientific thinking and promote construction of valid and reliable knowledge; Osborne, 2014). We did not focus on evaluation as a mode of assessment or self-regulated learning.

Mere exposure to scientific information will not help students to think critically (Sinatra, Kienhues, & Hofer, 2014). Rather, students must engage in evaluative processes, such as those used to weigh the connections between evidence and explanations. To be critical, students should consider how specific lines of evidence support or refute alternative explanations about a particular phenomenon (McNeill, Lizotte, Krajcik, & Marx, 2006; West, Toplak, & Stanovich, 2008). Critical evaluations may help students to gauge the more plausible of the alternatives (Lombardi, Nussbaum, & Sinatra, 2016). Doing so may not only increase students' critical thinking skills but also their understanding of scientific knowledge and how scientific knowledge is constructed. Evaluating alternative explanations and selecting the most plausible explanation based on evidentiary support increases cognitive engagement and elaboration and encourages construction of knowledge that is scientifically valid (Osborne, 2014).

One promising instructional scaffold that may promote critical evaluations about the connections between evidence and alterative explanations is the MEL diagram. Chinn and colleagues (Chinn & Buckland, 2012; Rinehart, Duncan, & Chinn, 2014) developed the original mode and structure of the MEL diagram for use in middle school life science classrooms. Lombardi et al. (2013) created a MEL diagram activity for the topic of climate change and used this MEL to facilitate shifts in middle school students' plausibility judgments and knowledge toward the scientifically accepted model of human-induced climate change. One of the main motivations for our project and present study was to examine whether more frequent engagement in evaluating connections between lines of evidence and alternative explanations consistently promoted reappraisal of plausibility about controversial, abstract, and/or complex Earth science topics. We also looked at the relations between evaluation, plausibility judgments, and knowledge about these Earth science concepts.

3 | METHODS

3.1 | Setting and participants

The present study examines data from the second year of a 3-year project involving high school (Grade 9–12) students from two different school systems in the United States. Students completed the measures and instructional activities during their Earth science classes, and teachers integrated these activities into their schools' scheduled curricula. Four teachers independently facilitated the activities; two teachers (referred to as SW1 and SW2) were located in an urban district in the Southwest United States and two (MA1 and MA2) were in suburban districts in the Mid-Atlantic United States. Each teacher had between two and six classes of participating students, for a total of 16 classes involved in the study. We received assent to conduct our research, as well as parental consent, from 339 participants. However, in a careful inspection of participant data, 40 gave patterned responses (e.g., answering questions in pattern that indicates participants did not read the questions, such as selecting straight 3s or selecting responses that result in a design such as a Christmas tree) that reflected disengagement (Gobert, Baker, & Wixon, 2015), and as such, were indicators of poor

TABLE 1 Race, ethnicity, and socioeconomic composition of the schools involved in the study

	School (%)			
Characteristic	**SW1[a]**	**SW2[a]**	**MA1[b]**	**MA2[b]**
American Indian/Alaskan native	0.80			
Asian	4.30	4.30	8.70	2.80
Black	17.1	11.9	1.50	7.70
Hispanic	23.9	63.1	3.40	10.0
Pacific islander	1.70		0.40	
Two or more races	7.30	4.10	0.60	1.80
Unknown race or ethnicity		2.30		
White	44.9	14.3	85.4	77.7
Free or reduced price lunch	33.2	72.1	2.20	27.4

Note. SW1 = data from Southwest Teacher 1's school, SW2 = data from Southwest Teacher 2's school, MA1 = data from Mid-Atlantic Teacher 1's school, and MA2 = data from Mid-Atlantic Teacher 2's school.
[a][SW] Department of Education (2016).
[b]State of [MA] Department of Education (2016).

data quality. We eliminated these participants from the study sample, with this disengagement distribution skewed slightly toward SW1 and SW2 (about 44% taught by SW1, $n = 18$; 33% taught by SW2, $n = 13$; 8% taught by MA1, $n = 3$; and 14% taught by MA2, $n = 6$). This made the final sample size $N = 299$, with about 34% taught by SW1, $n = 101$; 25% by SW2, $n = 76$; 16% by MA1, $n = 48$; and 25% by MA2, $n = 74$. Just over half ($n = 169$, 56.5%) of the participants were female. We did not collect race and ethnicity data from the participants, but the teachers indicated that participants reflected the general racial and ethnic composition that was characteristic of their schools (Table 1). In summary, the participants in this study represent a wide variety of demographic characteristics.

3.2 | Materials

The teachers collected data over the course of one full school year. Each class completed the MEL activities, described in more detail below, as an introduction to the respective instructional units.

3.2.1 | Instructional scaffolds

We adapted and applied the MEL diagram to four individual Earth science topics: climate change, fracking and earthquakes, wetlands and land use, and formation of Earth's Moon. We chose these topics because each has multiple plausible explanatory models (a scientifically accepted and an alternative) that students could evaluate. The four MEL topics (climate change, fracking, wetlands, and the Moon) represent a wide range of topics that might be covered in a typical high school Earth science scope and sequence. The MELs also presented lines of evidence that relate to each model (see Figure 1 for a student example of the climate change MEL activity). We briefly describe the topics of the four MEL activities below, with Table 2 summarizing the models and associated evidence statements. We also discuss the format of the MEL activities, as well as some preliminary research we have done on these activities in this section. However, for more details about the development of these activities, including alignment with the high school Earth science NGSS and instructional use guidance, see Bailey, Girtain, and Lombardi (2016), Holzer, Lombardi, and Bailey (2016), Hopkins, Crones, Burrell, Bailey, and Lombardi (2016), and Lombardi (2016).

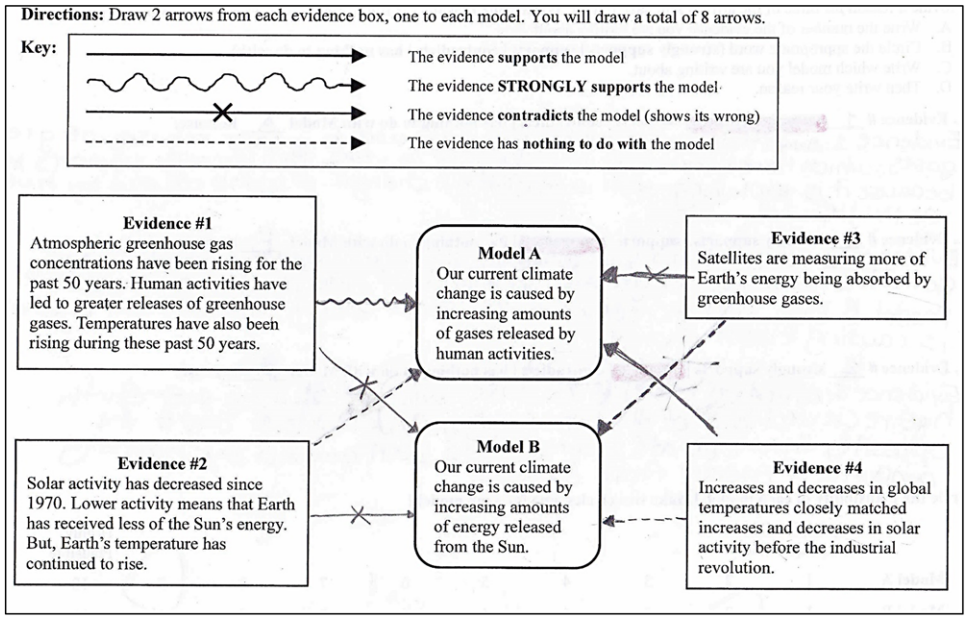
Climate change MEL

Climate change is a particularly relevant issue in today's society, and a topic that is already testing students' and teachers' ability to understand plausibility of scientific explanations (see, e.g., Lombardi, Danielson, & Young, 2016; Lombardi et al., 2013; Lombardi, Seyranian, & Sinatra, 2014; Lombardi & Sinatra, 2012, 2013). As a result, students may enter the

TABLE 2 Summary of models and evidence statements for each MEL activity

	Model		
Topic	**Scientific**	**Alternative**	**Evidence statements**
Climate change	Our current climate change is caused by increasing amounts of gases released by human activities. (Model A)	Our current climate change is caused by increasing amounts of energy released from the Sun. (Model B)	#1: Atmospheric greenhouse gas concentrations have been rising for the past 50 years. Human activities have led to greater releases of greenhouse gases. Temperatures have also been rising during these past 50 years.
			#2: Solar activity has decreased since 1970. Lower activity means that Earth has received less of the Sun's energy. But, Earth's temperature has continued to rise.
			#3: Satellites are measuring more of Earth's energy being absorbed by greenhouse gases.
			#4: Increases and decreases in global temperatures closely matched increases and decreases in solar activity before the industrial revolution.
Fracking	The increase in moderate magnitude earthquakes in the midwest is caused by fracking for fossil fuels. (Model A)	The increase in moderate magnitude earthquakes in the midwest is caused by normal tectonic plate motion. (Model B)	#1: Fracking fills cracks in the ground with water reducing friction between parts of the Earth's crust.
			#2: During the last 4 years, the number of earthquakes in the middle of the United States was 11 times higher than the 30-year average.
			#3: Convection of hot but solid and ductile rocks in the upper mantle creates stresses in the Earth's crust. These stresses cause the Earth's crust to fracture.
			#4: During fracking, cracks created by drilling boreholes do not create enough force to shift Earth's crustal plates.
Wetlands	Wetlands provide ecosystem services that contribute to human welfare and help sustain the biosphere.[a] (Model A)	Wetlands are a nuisance to humans and provide little overall environmental benefit. (Model B)	#1: Wetlands play a role in the global cycles of carbon, nitrogen, and sulfur. Wetlands change these nutrients into different forms necessary to continue their global cycles.
			#2: Flooding is a natural occurrence in low-lying areas, and wetlands are places where floodwaters can collect.
			#3: Wetlands contribute 70% of global atmospheric methane from natural sources.
			#4: Many wetlands are located in rapidly developing areas of the country.
Moon	The Moon formed after a large object collided with Earth and material from both combined to create the Moon. (Model B)	The Moon was an object that came from elsewhere in the solar system and was captured by Earth's gravity. (Model A)	#1: Earth's average density is higher than the Moon's. The density of Earth's crust is a little less than the Moon's, but Earth's density increases toward the core.
			#2: Simulations of other star systems show that planets form when smaller objects collide.
			#3: The Moon's orbit around Earth is tilted compared to the planets' orbits around the Sun.
			#4: Earth is about 35% iron, most of which is in the core. The Moon has very little iron.

[a]Although a socioscientific topic, the wetlands MEL asks students to make judgments about "value" models rather than scientific explanatory models. The "scientific" model in this case is that with which most environmental scientists agree.



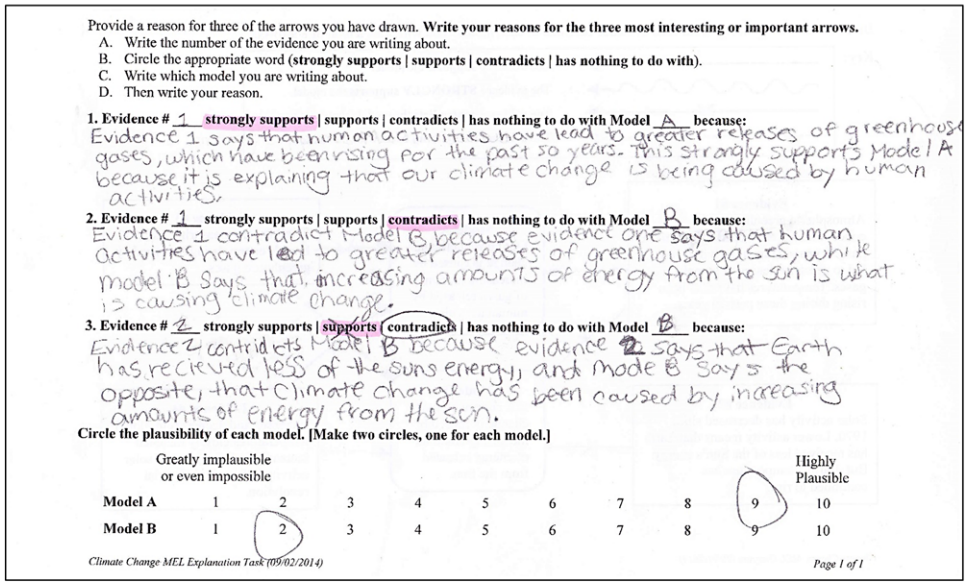


FIGURE 1 A student example of the climate change MEL diagram—top—with explanatory tasks on the bottom [Color figure can be viewed at wileyonlinelibrary.com]

classroom with very strong beliefs about the cause of climate change, and these may or may not be based in scientific understanding (Leiserowitz & Smith, 2010). We selected two plausible explanations about the cause of current climate change for this MEL, the scientifically accepted explanation that humans are the cause of current climate change (Cook et al., 2016), and the alternative explanation that current climate change is caused by an increased amount of solar irradiance (Boyes & Stanisstreet, 1993; Pruneau, Gravel, Bourque, & Langis, 2003; Shepardson, Choi, Niyogi, & Charusombat, 2011). The four lines of evidence in the climate change MEL cover atmospheric emissions and concentrations of greenhouse gases, recent changes in solar activity, observed effects of greenhouse gases on Earth's energy budget, and past correlations between solar activity and global temperatures. Lombardi et al. (2013) discuss the development, use, and validity of the climate change MEL used in a study with middle-grades students.

Fracking and earthquakes MEL

Recent increases of hydraulic fracturing, or fracking, is responsible for increases in fossil fuel production in the United States. Scientists have also recently associated this increased fossil fuel production with an increased occurrence of moderate-magnitude earthquakes in the midwest. Although some scientists agree that this represents a causal connection with fracking activities (Petersen et al., 2016), others have attributed increases in midwestern earthquakes to natural adjustments of a plate boundary (Oskin, 2015). Therefore, we aligned the two models for the fracking MEL with these two alternative explanations. The lines of evidence presented on this MEL include information about both the processes behind fracking and the occurrence of earthquakes: the effect that fracking injection has on friction in Earth's crust, the changes in frequency of earthquakes near fracking sites, the natural processes that cause earthquakes, and the amount of force exerted on Earth's crust during fracking.

Wetlands and land use MEL

Unlike the other activities, the wetlands MEL presents scientific context for value judgments rather than scientific processes. We developed the wetlands MEL with the idea that understanding of many Earth science topics can have immediate effects on our everyday lives, and, therefore, students should have practice applying scientific literacy both to develop understanding and to make sense of its relevance. Groups have debated the value of wetlands, with one side wishing to preserve them for their environmental benefits and the other side wishing to use available land in other ways (U.S. Environmental Protection Agency, 2017). These two value judgment alternatives served as the explanatory models in our wetlands MEL. The accompanying lines of evidence discussed the role wetlands play in global cycles of carbon, nitrogen, and sulfur; the effect that wetlands have on flooding in low-lying areas; the contribution of global atmospheric methane from wetlands; and the location of wetlands compared to developing areas. Even though participants were not making judgments about a scientific process in this MEL, they needed to evaluate relevant scientific information to make judgments about the connections between lines of scientific evidence and explanations. We acknowledge that individuals may look at other types of evidence (e.g., economic benefits and costs) when evaluating such explanations about wetlands, but for the purpose of comparison, we kept the wetlands MEL focused on scientific lines of evidence.

Moon formation MEL

There are multiple hypotheses about the formation of Earth's Moon that have been seriously considered, each presenting different ideas about how much of the Moon's material came from the Earth and how the material ended up in orbit around us. Scientists now generally agree that the most likely case is the one presented by the *giant impact theory* (Hartmann & Davis, 1975), which states that a large object crashed into Earth and material from each formed into the Moon. This represents the scientifically accepted explanatory model in the Moon MEL. For the other, we chose to present a plausible alternative based on a historical explanation, called the *capture theory*, by which the Moon was an independent object captured by Earth's gravity (Clery, 2013). The Moon MEL presented participants with lines of evidence that address main aspects of Earth and the Moon that have driven scientists' understanding of the Moon's origin: the average densities of Earth and the Moon, the results of simulations of the formation of other planets, the characteristics of the Moon's orbit, and the presence of iron in both Earth and the Moon. The Moon MEL differs somewhat from the other three because it covers a complex topic that students may not easily relate to their daily lives.

MEL activity part 1

The first page of the MEL activity presented participants with the two explanatory models about a particular phenomenon (Figure 1). Although each MEL included one model which is scientifically accepted and one which is a plausible alternative, the MEL introduced each as "Model A" or "Model B" without indication of the validity of either. For each MEL, these two alternative explanatory models were surrounded by four numbered evidence statements, each of which had a corresponding "evidence text" that is about a page in length. The evidence texts included diagrams and tables to elaborate on the evidence statements, and teachers encouraged participants to use these one-page texts in completing the activity. On the first page of each MEL, participants drew arrows of different types between each

evidence statement and each alternative model. These types of arrows indicated participants' judgments about how well a line of evidence supported a model, where (a) a straight line arrow meant that the evidence supported a model, (b) a squiggly line arrow meant the evidence strongly supported a model, (c) a dotted line arrow meant the evidence had nothing to do with the model, and (d) a straight line arrow with an "X" through it meant that the evidence contradicted a model. Participants drew a total of eight arrows to construct each MEL diagram (Figure 1).

MEL activity part 2

The second page of the MEL activity, also referred to as the "explanation task," prompted participants to give explanations of either two or three of the eight links that they drew on the first page diagram (Figure 1). The purpose of this page was to facilitate students' constructions of written explanations that we later scored (i.e., we scored only the explanations, not the diagrams). The explanation task asked participants to describe links they consider important or interesting. Using a sentence prompt for each explanation, participants wrote down the model and evidence number that they chose to discuss, as well as the evidence to model connection strength they drew on the diagram (i.e., strongly supports, supports, contradicts, or has nothing to do with). This preface served as the beginning of participants' written explanations, next prompting evaluation with the word "because." For example, one participant's written explanation from the climate change MEL read, "[Evidence 1 strongly supports Model A because…] It showed the direct correlation with the CO_2 emissions and the temperature" (note: the section in brackets is part of a sentence frame given to students, with underlined portions filled in by students).

Our scoring of the MEL activity focused on evaluations participants wrote in the explanation task. We used a system for scoring explanations that Lombardi, Brandt, Bickel, and Burg (2016) developed using a qualitative content analysis from a previous study, which involved middle-grade students' written responses about the climate change MEL used in this study. Lombardi, Brandt et al. (2016) developed four categories of explanations that drew from the frameworks of both Driver, Leach, Millar, and Scott (1996) and Dole and Sinatra (1998). The categories established well-defined levels of evaluation to represent the accuracy and elaboration present in participants' responses.

The following highlights these four different types of evaluations that primarily reflect epistemic quality (e.g., analysis about the strength of the connections between lines of evidence and explanations), but also include related conceptual understandings; readers should consult Lombardi, Brandt et al. (2016) for more details. The first category of participants' evaluations, called *erroneous evaluations*, described written responses that represented an incorrect determination about a link. Participants who made an erroneous evaluation demonstrated an inability to make a legitimate connection between a line of evidence and model, perhaps from a lack of attention or understanding. Erroneous evaluations prevent deeper comprehension and evaluation from occurring. For example, one student claimed that fracking evidence #2 has nothing to do with Model A because, "Evidence provided is irrelevant to the model," which is an incorrect evaluation based on lack of conceptual understanding (see Table 2 for models and evidence statements in the fracking MEL). We generally categorized participant explanations that discussed inaccurate links as erroneous, aside from clearly more advanced answers such as conscientious use of elimination-based logic. The second category, *descriptive evaluations*, represented weak and/or trivial written explanations. These weak explanations were generally from superficial evaluations between a line of evidence and a model. One student wrote that fracking evidence #2 has nothing to do with Model B because "they are talking about two different [things]." Although such evaluations were not necessarily inaccurate, they reflect little thinking and reasoning about the epistemic quality of the connection.

The third category, *relational evaluations*, represented correct links with somewhat deeper understanding, but participants' written explanations failed to differentiate between lines of evidence and explanatory models. For example, one student explained that fracking evidence #2 strongly supports Model A because "it gives numerical data that proves an increase in earthquakes around fracking sites." In this case, the student displays conceptual understanding about the evidence. However, the written explanation provides little insight into the epistemic level of quality applied in connecting the line of evidence to the explanation. The fourth and final category, *critical evaluations*, represented the greatest level of explanation development. Within this category, participants demonstrated an understanding of the scientific concepts and were able to critique the links using scientific reasoning and an accurate representation

of the role evidence plays in judging model validity. With these types of responses, students also demonstrated an explicit understanding of the epistemic quality of their connection between a line of evidence and an explanation. One student wrote that fracking evidence #2 strongly supports model A because, "it illustrated how the number of earthquakes near fracking sites is far too high above the normal average to be the result of normal plate movement." Although the student does not specifically make a claim about the plausibility of the explanation in this response, the student is explicitly addressing the strength of the connection between the line of evidence and the explanation in way that evaluates the link's epistemic quality.

These four categories served as distinct levels of evaluation for numerically scoring each explanation (1 = erroneous, 2 = descriptive, 3 = relational, 4 = critical), allowing us to consider participants' written explanations quantitatively. Initially, the second and fourth authors independently scored each participant's explanations using these four categories and Lombardi, Brandt et al.'s (2016) rubric as a guide. Initial rater scores were at an acceptable level of agreement (interclass coefficient, ICC = .612). The raters met to reconcile discrepancies in scoring and came to unanimous agreement on explanation task scores. We used these agreed upon scores in the subsequent analysis.

3.2.2 | Judgments of model plausibility

For each of the four MEL activities, students recorded their plausibility judgments of each model pre- and postinstruction. Students gauged the plausibility of each model using a 1–10 scale (1 = greatly implausible and 10 = highly plausible), based on previous measures used by Lombardi and colleagues (Lombardi, Danielson, & Young, 2016; Lombardi et al., 2013). We calculated plausibility scores as ratings for the scientific model minus ratings for the alternative model (Table 2). A positive score indicated that a participant judged the plausibility of the scientific model as greater than the alternative model, a negative rating indicated belief that the alternative model was more plausible, and a value of zero indicated belief that both models were equally plausible.

3.2.3 | Knowledge

We created short, five-item knowledge instruments for each topic (climate change, fracking, wetlands, and the Moon), which participants completed both prior to and just after engaging in a specific MEL activity. Per the methods used in previous work, students rated each item on a 5-point Likert scale (1 = strongly disagree and 5 = strongly agree) indicating how closely scientists would agree with the statement (Lombardi, Danielson, & Young, 2016; Lombardi et al., 2013, 2014; Lombardi & Sinatra, 2013). In this way, answers reflected students' understandings about the related scientific processes rather than their personal beliefs or opinions on the topic. We developed these statements from information on which there is clear scientific consensus. When the instruments were first created, there were approximately 30 such questions for each of the topics. However, as is the case with many classroom-based research studies, feedback from the teachers and students necessitated changes in the study methods. Specifically, we reduced the length for all topics to only five items because teachers said that they were spending too much instructional time on survey administration and students were viewing these longer instruments as unit tests, which was not our intention. In reducing the length of the knowledge instruments, at least one question addressed each evidence statement. Therefore, we ended up using a total of 20 knowledge items over the course of the study (see the Appendix for a complete list of these items).

For the present study, the reliability of all 20 knowledge items was marginal at both pre- (α = .555) and postinstruction (α = .571). This is no surprise given that we reduced the length of the original knowledge measures due to instructional concerns. Therefore, we used the Spearman–Brown prediction formula (Brown, 1910; Spearman, 1910) to estimate the reliability of a larger number of equally reliable items. We found that extending this 20-item measure by just four more items (one more item per topic) would increase pre and post α-values to over .9, which suggests that the existing items reliably measure knowledge of the topics. Furthermore, our previous studies using the 27-item climate change knowledge instrument, from which we drew the five items for the present study, has had consistently good to very good reliability (Lombardi, Danielson, & Young, 2016; Lombardi et al., 2013, 2014; Lombardi & Sinatra, 2013).[2]

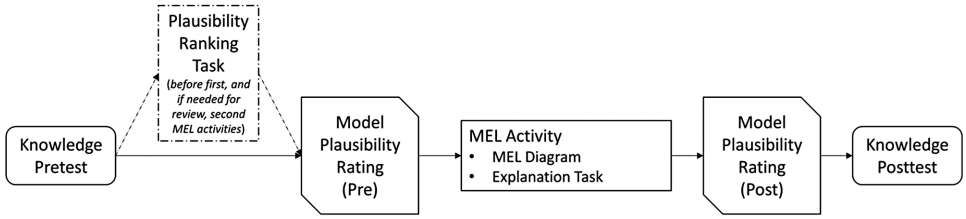


FIGURE 2 The sequence of tasks for a given MEL activity (i.e., climate change, fracking, wetlands, or Moon)

Finally, the advisory panel to our project, which consists of two Earth scientists and two educational psychologists, verified the face and content validity of our items.

3.3 | Procedures

Prior to conducting the year-long study, the research and development team, including the four master teachers whose students were involved in the study, met for 3 days of professional development focused on reflection, sharing, planning, and implementation. The reflection and sharing involved teacher-led discussions about how to best implement the instructional activities based on the pilot testing done in Year 1 of the project, which in turn influenced how they implemented instruction for this study. During the professional development, the teachers decided to introduce each MEL activity at the beginning of a unit prior to any instruction about the topic. The teachers and research team also decided as a group to avoid direct instruction in any phase of the MEL activities. In other words, the teachers presented the MEL activities using the instructions present on the handed out materials. Class discussions that arose only centered on elaboration and clarification of these instructions.

Students completed all of the activities over the course of a single school year, which included the full MEL activities and additional measures before and after all four MELs were completed. A breakdown of activities completed by students and the order in which they were implemented for each MEL is provided in Figure 2. Near the beginning of the year, prior to completing any MEL, students performed the "plausibility ranking task" as an introduction to the ideas of plausibility and critical evaluation. For this task, students were asked to rank the importance of different types of evidence for determining the plausibility of a model. These four types of evidence were the same as the links that students later indicated on the MEL: evidence that supports the model, strongly supports it, contradicts it, or has nothing to do with it. After ranking the importance of each from 1 to 4, they read a small passage on falsifiability that states scientific ideas cannot be proven but are rather disproven through opposing evidence and were then asked to rank the types of evidence again. This provided an introduction to the idea of plausibility for students and an initial look at the comfort with which they can evaluate the roles of scientific evidence. Teachers had the option of repeating or discussing this activity as a review prior to doing the second MEL if they felt it was needed.

Students completed each MEL at the beginning of an instructional unit related to the topic (e.g., the climate change MEL was the initial activity conducted in a unit on climate and weather, prior to any other instruction on the topic). For a given MEL, students began by completing the associated knowledge test, if needed the plausibility ranking task described above, and model plausibility ratings for that MEL. At this time, teachers also engaged the class in an unscripted short discussion about the models and the idea of plausibility, to clarify misunderstandings about the MEL process or address general questions about the topic. Students then began the MEL diagram, for which they were able to read the evidence texts and complete the first page of the MEL in groups. They then worked individually to write explanations on the second page of the MEL. Each MEL activity ended with the second iteration of the model plausibility ratings and knowledge test for that MEL. Upon completion of this sequence, teachers moved on to teaching their regular instructional unit.

Each implementation of a MEL activity (i.e., the sequence discussed in the paragraph above) took place over about two regular class periods (~90 minutes total). The teachers implemented the MEL activities during regular class time concurrently with their own planned curricula and at times when the topic of each MEL corresponded to scheduled

lessons. As a result, the timing and order in which students completed the MEL was different based on the teachers' discretion. Finally, each teacher reported that students enjoyed completing the MEL activities throughout the school year.

4 | RESULTS

We present the results in two sections. The first section addresses Research Question 1 (How do plausibility judgments and knowledge change when engaging in an instructional activity that facilitates high school students to think critically about controversial and/or complex Earth science topics?) and represents a fine-grained analysis examining pre- and post changes in knowledge and plausibility scores by participants' teacher groups. We recognized that potential differences in teaching, as wells as potential differences in topic, may influence outcomes, and we were curious about the robustness of the MEL activities to these variations. The second section addresses Research Question 2 (What are the relations between evaluation, plausibility reappraisal, and postinstructional knowledge, above and beyond background knowledge?) and represents a much broader survey of the overall relations between participants' evaluations, plausibility judgments, and knowledge.

4.1 | Research question 1: Changes in plausibility judgments and knowledge

We conducted two repeated measures multivariate analyses of variance (MANOVA) to examine our first research question. Prior to conducting the MANOVAs, we screened the data to ascertain alignment with assumptions inherent in ordinary least-squares analyses (OLS; e.g., MANOVA) about the normality and linearity of the sample, as well as assumptions about the equality of the homogeneity of variance–covariance matrices. We found four participants who consistently had *z*-scores less than −3 in many of the measured variables (plausibility, pre and post, for each MEL activity; knowledge, pre and post, for each MEL activity). After removing these four participants as univariate outliers, examination of Mahalanobis distance indicated no further outlier concerns. Almost all of the variables had skewness and kurtosis of absolute value less than or equal to 1, which some researchers use as general rule of thumb to indicate normality of the sample distribution (Nussbaum, 2014). The two exceptions were climate change knowledge, pre and post. Because the selection criteria that the absolute value of skewness and kurtosis should be less than or equal to 1 is arbitrary (Nussbaum, 2014), and given the robustness of OLS analyses to violations of normality (Osborne & Waters, 2002), we decided to retain all variables in the MANOVAs. Our examination of scatterplots for pair combinations of the measured variables also did not reveal any concerns with linearity. Finally, variance–covariance matrices were equivalent for both knowledge (Box's $M = 129$, $p = .255$) and plausibility (Box's $M = 117$, $p = .569$). The following subsections present the MANOVA results separately for the plausibility judgment and knowledge variables.

4.1.1 | Plausibility judgment

Table 3 shows the means for the plausibility judgment scores, with scores further broken down by MEL topic and participants' teacher group (e.g., SW1 = the group of students taught by Southwest Teacher 1). For the MANOVA, we only retained participants who completed all four plausibility measures, pre and post ($N = 153$). In this MANOVA, participants' teacher group was the between-subjects variable, time (pre and post) was the within-subjects variable, and climate change, fracking, wetlands, and Moon plausibility judgment scores were the dependent variables. The repeated measures MANOVA did not reveal a significant interaction between group and time for the combined plausibility judgment scores of all four topics, with Pillai's trace = .124, $F(12,444) = 1.59$, $p = .091$. However, there was a significant main effect in the combined plausibility judgment scores over time, with Pillai's trace = .298, $F(12,444) = 12.1$, $p < .001$, $\eta_p^2 = .099$ (moderate effect size; Tabachnick & Fidell, 2007).

Follow-up univariate ANOVAs showed there was a significant main effect over time for all groups, but only two topics, which both showed positive shifts toward the scientific model: climate change [mean shift = 1.60, $F(1,49) = 29.2$, $p < .001$, $\eta_p^2 = .164$, large effect size] and fracking [mean shift = 1.42, $F(1,149) = 22.7$, $p < .001$, $\eta_p^2 = .132$, large effect size]. However, there was not a significant shift in plausibility judgments for either wetlands [mean shift = −.281, $F(1,149) = .522$, $p = .471$] or Moon [mean shift = −.360, $F(1,149) = .749$, $p = .388$].

TABLE 3 Plausibility scores for each topic and participants' teacher group

Participants' teacher group	*n*	Preinstruction		Postinstruction		Shift	*p* value
		M	*SE*	*M*	*SE*		
				Climate			
SW1	36	0.25	0.49	0.50	0.46	0.25	.688
SW2	51	0.098	0.41	1.71	0.38	1.61	**.002**
MA1	29	0.97	0.54	3.86	0.51	2.90	**<.001**
MA2	37	0.54	0.48	2.43	0.45	1.89	**.002**
				Fracking			
SW1	36	−1.58	0.52	−0.83	0.57	0.75	.241
SW2	51	−1.92	0.43	−0.86	0.48	1.06	**.050**
MA1	29	−1.04	0.58	1.41	0.64	2.45	**.001**
MA2	37	−1.84	0.51	−0.081	0.56	1.76	**.006**
				Wetlands			
SW1	36	1.97	0.46	2.00	0.60	0.028	.965
SW2	51	2.10	0.38	1.73	0.50	−0.37	.486
MA1	29	2.66	0.51	3.17	0.66	0.52	.466
MA2	37	4.95	0.45	3.87	0.59	−1.08	.086
				Moon			
SW1	36	0.19	0.58	0.028	0.58	−0.17	.776
SW2	51	2.16	0.49	1.00	0.49	−1.16	**.020**
MA1	29	0.62	0.65	0.86	0.65	0.24	.711
MA2	37	1.70	0.57	1.78	0.57	0.081	.888

Notes. SW1 = the group of students taught by Southwest Teacher 1, SW2 = the group of students taught by Southwest Teacher 2, MA1 = the group of students taught by Mid-Atlantic Teacher 1, and MA2 = the group of students taught by Mid-Atlantic Teacher 2. For each topic, the possible range of plausibility scores is −9 to 9. Bolded values represent significant shifts in plausibility ($p \leq .05$).

Follow-up simple effects analyses revealed more nuanced results for each of the participants' teacher groups. For example, although there were no differences in groups in climate change preinstruction plausibility scores, both MA1 and MA2 showed appreciable shifts and had significantly higher postinstruction plausibility scores than SW1 (all *p*-values $\leq .018$). There were also no differences in groups in fracking preinstruction plausibility scores, but MA1 showed an appreciable shift and had significantly higher postinstruction plausibility scores than SW2 ($p = .030$). MA2 had significantly higher wetlands preinstruction plausibility scores when compared to the other groups (all *p*-values $\leq .006$). But, wetlands postinstruction plausibility scores increased appreciably for SW1 and MA1, and therefore there was no significant difference with MA2 after instruction (both *p*-values $\geq .165$). Finally, although SW2 experienced a significant decline in Moon plausibility scores from pre- to postinstruction, there was no significant difference between any of the groups at either pre- or postinstruction. Figure 3 shows the simple effects (as graphical columns) and main effects (as numerical values) for plausibility scores.

In terms of practical significance, effect sizes of the significant shifts range from medium to large, and overall represent shifts toward the scientifically accepted model. As a way of interpreting the results, significant shifts range from about one to three categories on a 10-category plausibility scale.

4.1.2. | Knowledge

Table 4 shows the means and detailed MANOVA results for knowledge scores, respectively, with scores broken down by MEL topic and participants' teacher group (e.g., SW1 = the group of students taught by Southwest Teacher 1).

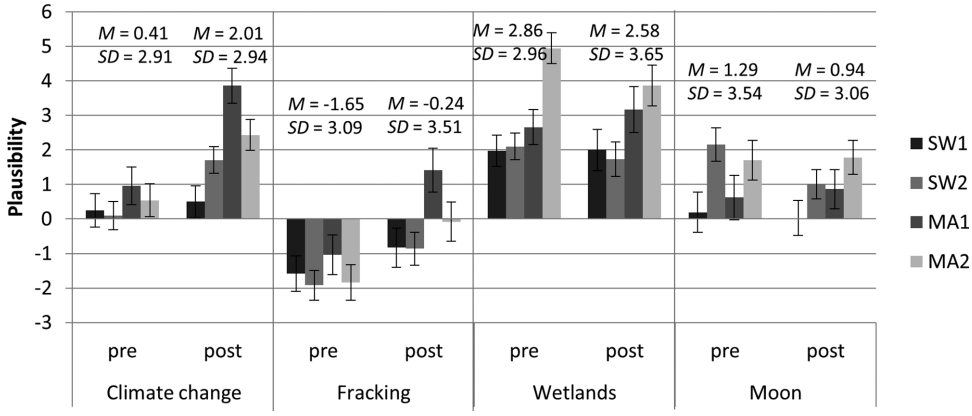


FIGURE 3 Model plausibility perception scores at pre- and postinstruction, with main effect means and standard deviations for each topic indicated numerically above the each set of columns, and simple effects by location indicated by individual column values. Bars on each column indicate ±1 standard error. The maximum score range was −9 to +9. SW1 = the group of students taught by Southwest Teacher 1, SW2 = the group of students taught by Southwest Teacher 2, MA1 = the group of students taught by Mid-Atlantic Teacher 1, and MA2 = the group of students taught by Mid-Atlantic Teacher 2

TABLE 4 Knowledge scores for each topic and participants' teacher group

		Preinstruction		Postinstruction			
Participants' teacher group	*n*	*M*	*SE*	*M*	*SE*	Gain	*p* value
				Climate			
SW1	40	18.8	0.46	20.5	0.38	1.68	**<.001**
SW2	59	19.3	0.36	19.5	0.31	0.21	.586
MA1	26	21.6	0.54	22.3	0.47	0.58	.315
MA2	62	19.4	0.35	20.3	0.30	0.88	**.018**
				Fracking			
SW1	40	16.2	0.36	18.0	0.42	1.80	**<.001**
SW2	59	15.8	0.30	17.3	0.35	1.46	**<.001**
MA1	26	17.7	0.44	18.5	0.52	0.77	.159
MA2	62	17.4	0.29	18.2	0.34	0.80	**.024**
				Wetlands			
SW1	40	15.3	0.40	17.7	0.45	2.38	**<.001**
SW2	59	16.5	0.33	17.5	0.37	1.00	**.018**
MA1	26	17.8	0.49	16.7	0.56	−1.08	.089
MA2	62	17.8	0.32	18.0	0.37	0.22	.595
				Moon			
SW1	40	18.2	0.36	19.1	0.46	0.93	**.040**
SW2	59	17.8	0.30	18.8	0.38	1.03	**.006**
MA1	26	19.7	0.45	20.2	0.58	0.54	.333
MA2	62	19.5	0.29	20.8	0.37	1.26	**.001**

Notes. SW1 = the group of students taught by Southwest Teacher 1, SW2 = the group of students taught by Southwest Teacher 2, MA1 = the group of students taught by Mid-Atlantic Teacher 1, and MA2 = the group of students taught by Mid-Atlantic Teacher 2. For each topic, the possible range of knowledge scores is 5–25. Values in bold represent significant knowledge gains ($p \leq .05$).

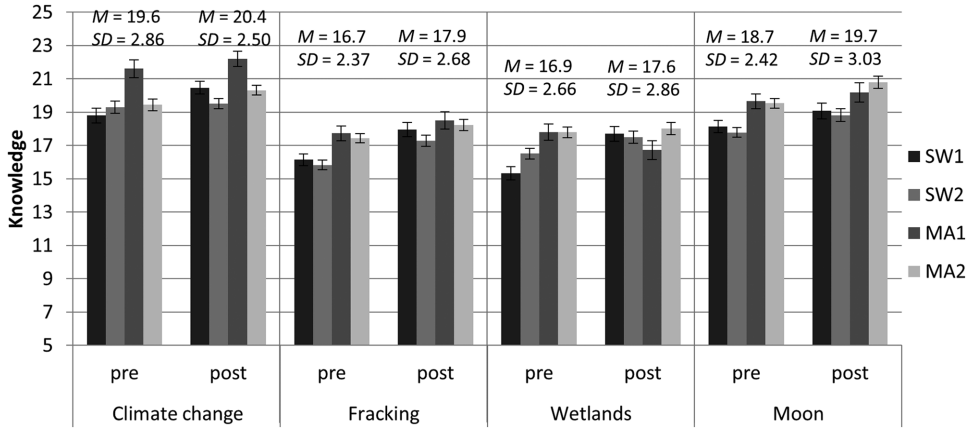


FIGURE 4 Knowledge scores at pre- and postinstruction, with main effect means and standard deviations for each topic indicated numerically above the each set of columns, and simple effects by location indicated by individual column values. Bars on each column indicate ±1 standard error. The maximum score range was 5–25. SW1 = the group of students taught by Southwest Teacher 1, SW2 = the group of students taught by Southwest Teacher 2, MA1 = the group of students taught by Mid-Atlantic Teacher 1, and MA2 = the group of students taught by Mid-Atlantic Teacher 2

A few more participants completed all four knowledge measures, pre and post ($N = 187$). In this MANOVA, participants' teacher group was the between-subjects variable, time (pre and post) was the within-subjects variable, and climate change, fracking, wetlands, and Moon knowledge scores were the dependent variables. Unlike plausibility judgment scores, the repeated measures MANOVA revealed a significant interaction between group and time for the combined knowledge scores of all four topics (climate, fracking, wetlands, and the Moon), with Pillai's trace = .159, $F(12,546) = 2.55$, $p = .003$, $\eta_p^2 = .053$ (small to medium effect size). There was also a significant main effect in the combined knowledge scores over time, with Pillai's trace = .251, $F(12,546) = 15.1$, $p < .001$, $\eta_p^2 = .251$ (large effect size).

Follow-up univariate ANOVAs showed there was a significant main effect over time for all topics and groups. Each topic showed a positive gain (i.e., postscores were always higher than prescores): climate change [mean gain = .797, $F(1,183) = 13.6$, $p < .001$, $\eta_p^2 = .069$, moderate effect size]; fracking [mean gain = 1.22, $F(1,183) = 31.4$, $p < .001$, $\eta_p^2 = .146$, large effect size]; wetlands [mean gain = .746, $F(1,183) = 6.34$, $p = .013$, $\eta_p^2 = .033$, small effect size]; and Moon [mean gain = 1.02, $F(1,183) = 18.2$, $p < .001$, $\eta_p^2 = .091$, moderate effect size].

Follow-up simple effects analyses showed specific differences for each of the participants' teacher groups. For example, both SW1 and SW2 had significantly lower fracking knowledge scores at preinstruction than MA1 and MA2 (all p-values $\leq .035$), but at postinstruction, there was no statistically significant difference in fracking knowledge scores between any of the participants' teacher groups (all p-values $\geq .097$). Somewhat similarly, SW1 and SW2 had lower wetlands knowledge scores than MA2 at preinstruction (both p-values $\leq .031$), but at postinstruction, there was no statistically significant difference in wetlands knowledge scores between any of the participants' teacher groups (all p-values $\geq .104$). In two instances, one group showed consistently higher scores at both pre- and postinstruction than the other three groups. First, MA1 had significantly higher pre- and postinstruction climate knowledge scores than the other three groups (all p-values $\leq .027$). Second, MA2 had significantly higher pre- and postinstruction Moon knowledge scores than both SW1 and SW2 (all p-values $\leq .026$). Figure 4 shows the simple effects (as graphical columns) and main effects (as numerical values) for knowledge scores.

In terms of practical significance, effect sizes range from small to large, depending upon topic and participants' teacher group. As a way of interpreting the results, knowledge gains generally ranged from 3 to 5% per topic, with some groups' knowledge gains as much as 10%. With each of the activities lasting only about 90 minutes total (two class periods of traditional instruction), such shifts may have practical significance when these activities are considered within the context of a full unit of instruction (about 10–15 class periods).

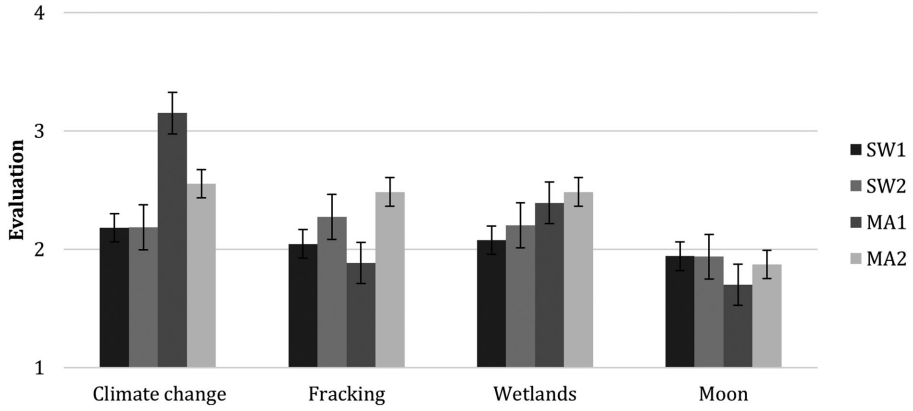


FIGURE 5 Evaluation scores by teacher and topic. Bars on each column indicate ±1 standard error. The score range was 1 to 4. SW1 = the group of students taught by Southwest Teacher 1, SW2 = the group of students taught by Southwest Teacher 2, MA1 = the group of students taught by Mid-Atlantic Teacher 1, and MA2 = the group of students taught by Mid-Atlantic Teacher 2

4.2 | Research question 2: Relations between evaluation, plausibility reappraisal, and knowledge

Figure 5 shows evaluation scores by teacher for each MEL activity. We conducted a MANOVA to analyze evaluation score differences, with teacher as the grouping variable and evaluation scores as the dependent variable. There was a significant difference between evaluations scores, with Pillai's trace = .413, $F(12,486) = 6.46$, $p < .0001$, $\eta_p^2 = .138$ (moderate effect size). However, a post hoc Tukey honest significant difference (HSD) analysis showed only two significant differences in evaluation scores (MA1 had significantly higher climate change evaluation scores than SW1, SW2, and MA2; and MA1 had significantly lower fracking evaluation scores than SW2 and MA2). In other words, almost all evaluation scores for the MEL activities were the same, ranging between a value of one (descriptive evaluations) and two (relational evaluations). Because of the relative uniformity in evaluation scores, our analysis for Research Question 2 represents a broad examination of the overall relation between evaluation, plausibility reappraisal, and knowledge through structural equation modeling (SEM).

The SEM analysis compares the direct and mediated relations between evaluation (as measured by the activities' explanation tasks), plausibility reappraisal (as the simple gain in plausibility judgments, i.e., the difference between post -and preinstruction plausibility scores), and postinstructional knowledge. In addition to comparing the direct and mediated relations between evaluation, plausibility reappraisal, and postinstructional knowledge, we also examined the direct relation between pre- and postinstructional knowledge to gauge the impact of the instructional activities above and beyond background (i.e., preinstructional) knowledge.

We used variance-based structural equation modeling (VB-SEM) to examine these various relational paths, and specifically used the Warp PLS v.4.0 statistical software (Kock, 2013). Unlike traditional, covariance-based structural equation modeling, which assumes that all data are metric/continuous and conform to a normal distribution, VB-SEM uses the partial least-squares method, which is based on ranked data and is distribution-free. Use of ranked-based data allows for more statistical power without compromising or inflating the chance for Type I errors for a large range of sample sizes and variation of group sizes (Reinartz, Haenlein, & Henseler, 2009). VB-SEM and partial least-squares methods have been used increasingly in social science research (Esposito Vinzi, Chin, Henseler, & Wang, 2010) and are being used more frequently in educational research (see, e.g., Hagger, Sultan, Hardcastle, & Chatzisarantis, 2015; Lombardi, Danielson, & Young, 2016).

We constructed the latent variables in our model (postinstructional plausibility reappraisal, preinstructional knowledge, postinstructional knowledge, and evaluation; Figure 6) using scores for all four topics (climate change, fracking, wetlands, and Moon), respectively. Furthermore, because this analysis represents a broad view of these relations (i.e.,

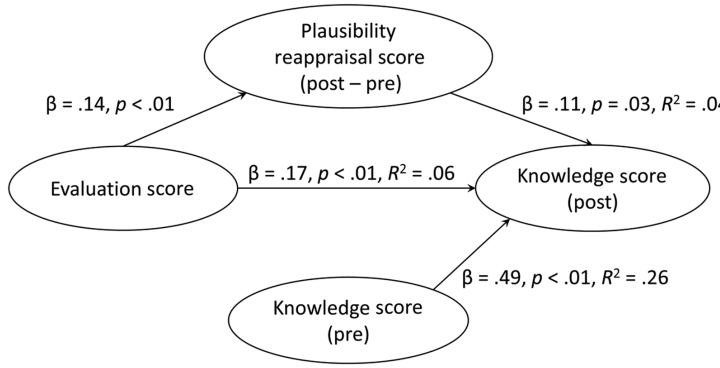


FIGURE 6 Comparison of mediated (by plausibility reappraisal) and direct causal models showing the relation of evaluation to postinstructional knowledge, above and beyond preinstructional (background) knowledge. Plausibility reappraisal score was calculated as the simple gain (i.e., the postinstructional plausibility score minus the preinstructional plausibility score)

as compared to the fine-grained analysis for Research Question 1, discussed above), we replaced any missing values using the arithmetic mean imputation method (i.e., replacing missing values with column averages; Kock, 2014). None of the variable columns had greater than 20% missing data and is therefore below the threshold where the imputation method introduces significant biases, from the perspective of model testing (Hair, Black, Babin, & Anderson, 2009). After excluding the four outliers who consistently had *z*-scores less than −3 in many of the measured variables, the sample size in the VB-SEM was $N = 295$. The overall means and standard deviations of the four latent variables were: evaluation ($M = 2.20$, $SD = 0.35$), plausibility reappraisal ($M = 0.66$, $SD = 3.82$), preinstructional knowledge ($M = 18.2$, $SD = 1.55$), and postinstructional knowledge ($M = 18.9$, $SD = 1.60$).

We used several fit and quality indices to gauge the validity of our variance-based structural equation model. These indices include overall goodness-of-fit (GoF), average path coefficient (APC), average coefficient of determination across the model (average R^2 or ARS), average variance inflation factor for model parameters (AVIF), and average full collinearity VIF (AFVIF). Tenenhaus, Amato, and Esposito Vinzi (2004) proposed that researchers use GoF as a criterion for the overall model prediction performance based on both the measurement and the structural model. A model has a large explanatory power when GoF is greater than 0.36 (Wetzels, Odekerken-Schroder, & van Oppen, 2009). Both APC and ARS provide further information about model adequacy and should have values that are statistically significantly different from zero (Hagger et al., 2015). Finally, high AVIF and AFVIF values indicate a potentially large degree of collinearity (i.e., redundancy of variables; Tabachnick & Fidell, 2007) is present in the model. Values of AVIF and AFVIF should generally be below 3.3 (Kock & Lynn, 2012) to ensure that variables are not redundant. For the present study, the overall fit and quality of model was good, with GoF = .244 (medium effect size; Tenenhaus, Esposito Vinzi, Chatelin, & Lauro, 2005); APC = .228, $p < .001$; ARS = .178, $p < .001$; AVIF = 1.07; and AFVIF = 1.26.

We included the model's standardized path values in Figure 6. We chose standardized values because this allows the reader to compare differences of magnitude between predictors with different scales. The direct path between evaluation and postinstructional knowledge was significant, with $\beta = .17$, $p < .01$. Furthermore, the mediated relations were also significant, with $\beta = .14$, $p < .01$, for the evaluation to plausibility reappraisal path, and $\beta = .11$, $p = .03$, for the plausibility reappraisal to knowledge path. In other words, higher evaluation and greater plausibility reappraisal were related to higher postinstruction knowledge scores. The total amount of variance in postinstructional knowledge explained by the model was 36%, with the direct pathway involving evaluation accounting for about 6% of the variance in postinstructional knowledge and the mediated pathway involving the relation between evaluation and plausibility reappraisal accounting for an additional 4% of the variance. In total, 10% of the variance in postinstructional knowledge was related to evaluation scores, which represents almost a 40% increase above and beyond background knowledge.

5 | DISCUSSION

The results from the present study look promising when considering both the fine-grained analysis we used to examine Research Question 1 and the broader analysis we used with Research Question 2. The MANOVA results show that some of the MELs promoted shifts in plausibility judgments toward scientifically accepted explanations, and overall knowledge increased, although modestly, in all the MELs. Furthermore, the structural equation model showed that higher levels of evaluation related to greater postinstructional knowledge, with part of this relation mediated by plausibility reappraisal. These results may support the theoretical position taken by Lombardi, Nussbaum, & Sinatra (2016) and the idea that evaluation is a process that is central to learning about science (Ford, 2015). However, as is often the case with research conducted in classroom settings, the results are not completely clear-cut. Therefore, we approach this general claim cautiously and recognize the need for more research in this area, with a potential reframing of the role of evaluation and plausibility in knowledge construction. In the case of two topics, wetlands and Moon formation, participants' plausibility judgments did not significantly shift after the instructional activities. For the wetlands and Moon MELs, other types of cognitive processing associated with evaluation (e.g., elaboration and organization that may not be directly related to plausibility judgments) may have contributed to postinstructional knowledge (Dole & Sinatra, 1998; Lombardi, Nussbaum, and Sinatra, 2016). This speculation warrants further research.

5.1 | Implications for instruction

Evaluative reasoning is foundational to many, if not all, scientific practices (Ford, 2015). Instruction that promotes evaluation may present an opportunity for students to more fully engage in scientific practices and deepen their understanding of core disciplinary content (NRC, 2012). Our findings here, along with previous research (see, e.g., Lombardi et al., 2013), serve as a potential indication for evaluation-based lessons to promote students' active reappraisal of scientific explanations. Such lessons could easily replace more passive lessons based on telling and transmission of knowledge from the teacher to the student (e.g., lectures, fill-in the blank worksheets), especially because the type of materials used in this study still involve reading of scientific text. The study results serve as a legitimate justification for students to construct their own scientific understanding through authentic evaluation and critique.

Encouraging the application of scientific evaluation and explicit consideration of plausibility judgments may be particularly important when students confront unresolved concepts. Such topics may necessitate that students actively construct their own understanding to make sense of the validity of more scientifically rooted ideas. Students should also recognize how the faults and facts of multiple perspectives allow for reasonable and informed evaluations. Critical evaluation may be facilitated a variety of classroom activities, such as the MEL diagram, critical questions and argument vee diagrams (Nussbaum & Edwards, 2011), metacognitive prompts (Peters & Kitsantas, 2010), openness to alternatives (Meyer & Lederman, 2013), peer evaluation of constructed explanations (Wang, 2015), self-regulation checklists (Peters, 2012), and discursive actions (e.g., judging the validity of differing views; Christodoulou & Osborne, 2014).

For any activity in which students evaluate the plausibility of alternative models, it is also important that teachers still clearly inform their students about which of the explanations is scientifically accepted. Some students' difficulty with evaluations could lead them toward lesser understanding. Therefore, we suggest following any activity comparing alternatives with further discussions of about the scientific validity of all of the connections between evidence and explanations. Most importantly, activities such as the MEL, where students consider alternative explanations, should not be thought of as "Teaching the Controversy" (i.e., a campaign started to elevate nonscientific viewpoints in the science classroom in a way that legitimizes mythological thinking; Foran, 2014). The goal of presenting alternatives during activities that promote evaluation is to help develop students' abilities to weigh the merits of alternatives, so that students can ultimately make more critical evaluations and conclusions when facing scientific phenomena.

This study suggests that higher levels of evaluation are beneficial for students' learning. Without explicit instruction, students may have trouble basing their cognitive judgments about scientific knowledge on evidence. Constructing evidence-based explanations is a fundamental basis for many of the NGSS performance expectations (NGSS Lead

States, 2013). The explicit process of evaluating the connections between evidence and models may be one way to promote more scientific judgments and deeper understanding (Eastwood, Schlegel, & Cook, 2011). When doing so, students are also able to play a more active role in their own learning processes. This active learning could make activities more engaging and help students be more effective in reappraising their personal judgments about knowledge (Sinatra, Heddy, & Lombardi, 2015). Engaging in instructional activities based on evaluation may encourage students to critique their prior understanding, thus promoting a habit of using the scientific practices when confronted with science topics. Application of evaluation in the classroom is an engaging way for students to develop deeper understanding of scientific knowledge and of the process for constructing scientifiyc knowledge.

5.2 | Limitations and concluding thoughts

Classroom-based studies can have strong ecological validity, but the nature of conducting such studies in authentic settings means that certain limitations are inherent in these types of designs. For example, the study participants, although from two very different regions and four separate populations, are not necessarily a representative sample that one can generalize to other school settings. This limitation raises the potential need for future work on the relationships between evaluation, plausibility reappraisal, and knowledge with other populations and age levels.

This particular study was not a comparison between types of instruction, and therefore, we cannot be completely confident that the materials used were responsible for shifts in plausibility and changes in knowledge, as well as the connection between evaluation, plausibility reappraisal, and knowledge. Although the resulting trends were consistent and complementary between both the fine-grained analysis associated with Research Question 1 and the broader analysis with Research Question 2, we cannot say with utmost certainty that other factors did not come into play. For example, it may be that participants were motivated to meaningfully engage in the study's instructional activities because they were novel, specifically when compared to traditional, lecture-based instruction (Sinatra et al., 2015). Participants may have also been more motivated because the activities asked the participants to be autonomous in their process of evaluation and knowledge construction (Ryan & Deci, 2000). Anecdotal evidence from the teachers whose students participated in the study suggests that students greatly enjoyed the study's instructional activities. It is also possible, but somewhat unlikely based on historical evidence (see, e.g., Katz, Kaplan, & Gueta, 2009), that the instructional activities may have motivated students to seek out additional information about these topics outside of class, particularly for those situations when the instruction ran over one class period. Other factors besides motivation may have also been involved, but because we measured knowledge and plausibility judgments just prior to and immediately after engaging in the study's instructional activities, as well as examining evaluation as part of the instructional activity (i.e., in scoring the explanation tasks), we have some confidence that the results are reflective of the intervention. However, we do acknowledge the need for additional comparative studies, and our research efforts are ongoing in this regard.

We concede that reducing the length of the knowledge instruments warrants caution in interpreting the results. For example, it may be that shorter instruments are prone to prime students to respond in a particular way. However, the results from both the shorter climate change knowledge instrument used in the present study and the longer climate change knowledge instrument used in earlier studies (see, e.g., Lombardi et al., 2013; Lombardi, Nussbaum, and Sinatra, 2016) showed similar results. Reducing the length of the instruments also resulted in lower score reliability. Such reduction is likely to attenuate results at the ends of sample distributions. This would most likely dampen the pre- to postinstructional differences revealed in the fine-grained analyses that we conducted to investigate Research Question 1.

Helping students develop their critical thinking skills is a daunting task for most educators; therefore, instructional tools and methods that facilitate this important 21st century skill would be of great benefit to educators. In the science classroom, being critical involves making evaluations about the validity of explanations based on lines of evidence. With the many controversial and complex socioscientific issues found in an Earth science curriculum, such as climate change and fracking, being explicitly evaluative of the connection between lines of evidence and alternative explanations may help students figure out which of all plausible alternatives is the most scientific. Engaging students in critical

evaluations may deepen their understanding about these phenomena, as well as their understanding of the scientific practices used to construct valid scientific knowledge.

ACKNOWLEDGMENTS

The U.S. National Science Foundation (NSF), under Grant No. DRL-1316057, supported this research. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the NSF's views. We wish to express sincere thanks to the four teachers and numerous students who invited us into their classrooms.

NOTES

[1] Some could consider plausibility to mean a scientific discipline's or community's plausibility judgment about a theory's explanations. But for the present study, we specifically focused on students' judgments about the plausibility of explanations.

[2] We acknowledge opening potential concerns about lower reliability of the knowledge instruments when we reduced the number of items. However, lower reliability tends to attenuate results, particularly at the ends of the sample distribution, due to higher signal to noise ratios (Osterlind, 2010). As such, lower reliability most likely dampens differences in distribution samples (i.e., in the present study, lower reliability could dampen pre- to postinstruction differences; Carmines & Zeller, 1979). Specifically, the pre- to post instructional differences that we found would be more likely to have a greater effect size if we had used longer instruments, with subsequently higher Cronbach's alpha.

ORCID


Doug Lombardi http://orcid.org/0000-0002-4172-318X

REFERENCES

Allchin, D. (2011). Evaluating knowledge of the nature of (whole) science. *Science Education*, *95*(3), 518–542. https://doi.org/10.1002/sce.20432

Bailey, J. M., Girtain, C., & Lombardi, D. (2016). Understanding the formation of the Earth's Moon. *The Earth Scientist*, *32*(2), 11–16.

Bailin, S. (2002). Critical thinking and science education. *Science & Education*, *11*(4), 361–375. https://doi.org/10.1023/A:1016042608621

Boyes, E., & Stanisstreet, M. (1993). The "greenhouse effect": Children's perceptions of causes, consequences and cures. *International Journal of Science Education*, *15*(5), 531–552. https://doi.org/10.1080/0950069930150507

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, *95*(4), 639–669. https://doi.org/10.1002/sce.20449

Brewer, W. F., Chinn, C. A., & Samarapungavan, A. (1998). Explanation in scientists and children. *Minds and Machines*, *8*, 119–136. https://doi.org/10.1023/A:1008242619231

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Bybee, R. W. (2009). *The BSCS 5E instructional model and 21st century skills*. Washington, DC: National Academies Board on Science Education. Retrieved from https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_073327.pdf

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Chinn, C. A., & Buckland, L. A. (2012). Model-based instruction: Fostering change in evolutionary conceptions and in epistemic practices. In K. S. Rosengren, E. M. Evans, S. Brem, & G. M. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning about evolution* (pp. 211–232). New York, NY: Oxford University Press.

Christodoulou, A., & Osborne, J. (2014). The science classroom as a site of epistemic talk: A case study of a teacher's attempts to teach science based on argument. *Journal of Research in Science Teaching*, *51*(10), 1275–1300. https://doi.org/10.1002/tea.21166

Clery, D. (2013). Impact theory gets whacked. *Science*, *342*(6155), 183–185. https://doi.org/10.1126/science.342.6155.183

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R., Verheggen, B., Maibach, E. W., … & Nuccitelli, D. (2016). Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, *11*(4), 1–7. https://doi.org/10.1088/1748-9326/11/4/048002

Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist*, *33*(2), 109–128. https://doi.org/10.1080/00461520.1998.9653294

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher*, *23*(7), 5–12. https://doi.org/10.3102/0013189X023007005

Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science*. Buckingham, England: Open University Press.

Duschl, R. A. (2007). Quality argumentation and epistemic criteria. In S. Erduran and M. P. Jiménez-Aleixandre (Eds.), *Reconceptualizing the nature of science for science education* (pp. 159–175). Dordrecht, The Netherlands: Springer.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education, *Studies in Science Education*, *38*(1), 39–72. https://doi.org/10.1080/03057260208560187

Eastwood, J. L., Schlegel, W. M., & Cook, K. L. (2011). Effects of an interdisciplinary program on students' reasoning with socioscientific issues and perceptions of their learning experiences. In T. Sadler (Ed.), *Socio-scientific issues in the classroom* (pp. 89–126). Dordrecht, The Netherlands: Springer.

Esposito Vinzi, V., Chin, W. W., Henseler, J., & Wang, H. (2010). *Handbook of partial least squares: Concepts, methods, and application*. Berlin, Germany: Springer.

Foran, C. (2014). The plan to get climate-change denial into schools. *The Atlantic*. Retrieved from https://www.theatlantic.com/education/archive/2014/12/the-plan-to-get-climate-change-denial-into-schools/383540/

Ford, M. J. (2015). Educational implications of choosing "practice" to describe science in the Next Generation Science Standards. *Science Education*, *99*(6), 1041–1048. https://doi.org/10.1002/sce.21188

Giere, R. N. (2010). *Explaining science: A cognitive approach*. Chicago, IL: University of Chicago Press.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist*, *50*(1), 43–57. https://doi.org/10.1080/00461520.2014.999919

Hagger, M. S., Sultan, S., Hardcastle, S. J., & Chatzisarantis, N. L. (2015). Perceived autonomy support and autonomous motivation toward mathematics activities in educational and out-of-school contexts is related to mathematics homework behavior and attainment. *Contemporary Educational Psychology*, *41*, 111–123. https://doi.org/10.1016/j.cedpsych.2014.12.002

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice Hall.

Hartmann, W. K., & Davis, D. R. (1975). Satellite-sized planetesimals and lunar origin. *Icarus*, *24*(4), 504–515. https://doi.org/10.1016/0019-1035(75)90070-6

Henderson, J. B., MacPherson, A., Osborne, J., & Wild, A. (2015). Beyond construction: Five arguments for the role and value of critique in learning science. *International Journal of Science Education*, *37*(10), 1668–1697. https://doi.org/10.1080/09500693.2015.1043598

Hogan, K., & Maglienti, M. (2001). Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions. *Journal of Research in Science Teaching*, *38*(6), 663–687. https://doi.org/10.1002/tea.1025

Holzer, M. A., Lombardi, D., & Bailey, J. M. (2016). Wetlands: Good or bad? Evaluating competing models with a MEL diagram. *The Earth Scientist*, *32*(2), 17–22.

Hopkins, J. D., Crones, P., Burrell, S., Bailey, J. M., & Lombardi, D. (2016). Using the model evidence link (MEL) diagram to evaluate the connections between fracking and earthquakes. *The Earth Scientist*, *32*(2), 23–30.

Katz, I., Kaplan, A., & Gueta, G. (2009). Students' needs, teachers' support, and motivation for doing homework: A cross-sectional study. *The Journal of Experimental Education*, *78*(2), 246–267. https://doi.org/10.1080/00220970903292868

Kock, N. (2013). *WarpPLS 4.0 user manual*. Laredo, TX: ScriptWarp Systems.

Kock, N. (2014). *Single missing data imputation in PLS-SEM*. Laredo, TX: ScriptWarp Systems.

Kock, N., & Lynn, G. S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, *13*(7), 546–580.

Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development*, *1*(1), 113–129. https://doi.org/10.1207/S15327647JCD0101N_11

Lederman, N. G. (1992). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching*, *29*(4), 331–359. https://doi.org/10.1002/tea.3660290404

Leiserowitz, A., & Smith, N. (2010). *Knowledge of climate change across global warming's six Americas*. New Haven, CT: Yale University. Yale Project on Climate Change Communication.

Lombardi, D. (2016). Beyond the controversy: Instructional scaffolds to promote critical evaluation and understanding of Earth science. *The Earth Scientist*, *32*(2), 5–10.

Lombardi, D., Brandt, C. B., Bickel, E. S., & Burg, C. (2016). Students' evaluations about climate change. *International Journal of Science Education*, *38*(8), 1392–1414. https://doi.org/10.1080/09500693.2016.1193912

Lombardi, D., Danielson, R. W., & Young, N. (2016). A plausible connection: Models examining the relations between evaluation, plausibility, and the refutation text effect. *Learning and Instruction*, *44*, 74–86. https://doi.org/10.1016/j.learninstruc.2016.03.003

Lombardi, D., Nussbaum, E. M., & Sinatra, G. M. (2016). Plausibility judgments in conceptual change and epistemic cognition. *Educational Psychologist*, *51*(1), 35–56. https://doi.org/10.1080/00461520.2015.1113134

Lombardi, D., Seyranian, V., & Sinatra, G. M. (2014). Source effects and plausibility judgments when reading about climate change. *Discourse Processes*, *51*(1-2), 75–92. https://doi.org/10.1080/0163853X.2013.855049

Lombardi, D., & Sinatra, G. M. (2012). College students' perceptions about the plausibility of human-induced climate change. *Research in Science Education*, *42*(2), 201–217. https://doi.org/10.1007/s11165-010-9196-z

Lombardi, D., & Sinatra, G. M. (2013). Emotions about teaching about human-induced climate change. *International Journal of Science Education*, *35*(1), 167–191. https://doi.org/10.1080/09500693.2012.738372

Lombardi, D., Sinatra, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction*, *27*, 50–62. https://doi.org/10.1016/j.learninstruc.2013.03.001

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, *15*(2), 153–191. https://doi.org/10.1207/s15327809jls1502_1.

Meyer, A. A., & Lederman, N. G. (2013). Inventing creativity: An exploration of the pedagogy of ingenuity in science classrooms. *School Science and Mathematics*, *113*(8), 400–409. https://doi.org/10.1111/ssm.12039

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For states by states*. Washington, DC: The National Academies Press.

Nussbaum, E. M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist*, *46*(2), 84–106. https://doi.org/10.1080/00461520.2011.558816

Nussbaum, E. M. (2014). *Categorical and nonparametric data analysis*. New York, NY: Routledge.

Nussbaum, E. M., & Edwards, O. V. (2011). Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices. *Journal of the Learning Sciences*, *20*(3), 443–488. https://doi.org/10.1080/10508406.2011.564567

Osborne, J. (2014). Teaching critical thinking? New directions in science education. *School Science Review*, *352*, 53–62.

Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, *8*(2). Retrieved from http://pareonline.net/getvn.asp?v=8&n=2

Oskin, B. (2015). Hidden faults explain earthquakes in fracking zones. Retrieved from https://www.livescience.com/49785-oklahoma-earthquake-risk-hidden-faults.html

Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and applications of mental appraisal* (2nd ed). Boston, MA: Pearson Education.

Peters, E. E. (2012). Developing content knowledge in students through explicit teaching of the nature of science: Influences of goal setting and self-monitoring. *Science & Education*, *21*(6), 881–898. https://doi.org/10.1007/s11191-009-9219-1

Peters, E., & Kitsantas, A. (2010). The effect of nature of science metacognitive prompts on science students' content and nature of science knowledge, metacognition, and self-regulatory efficacy. *School Science and Mathematics*, *110*(8), 382–396. https://doi.org/10.1111/j.1949-8594.2010.00050.x

Petersen, M. D., Mueller, C. S., Moschetti, M. P., Hoover, S. M., Llenos, A. L., Ellsworth, W. L., … Rukstales, K. S. (2016). *2016* One-year seismic hazard forecast for the Central and Eastern United States from induced and natural earthquakes (No. 2016-1035). Reston, VA: U.S. Geological Survey.

Pintrich, P. R., Marx, R. W., & Boyle, R. B. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research*, *63*(2), 167–199. https://doi.org/10.3102/00346543063002167

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227. https://doi.org/10.1002/sce.3730660207

Pruneau, D., Gravel, H., Bourque, W., & Langis, J. (2003). Experimentation with a socio-constructivist process for climate change education. *Environmental Education Research*, *9*(4), 429–446. https://doi.org/10.1080/1350462032000126096

Reinartz, W. J., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance based SEM. *International Journal of Market Research*, *26*(4), 332–344. https://doi.org/10.1016/j.ijresmar.2009.08.001

Rinehart, R. W., Duncan, R. G., & Chinn, C. A. (2014). A scaffolding suite to support evidence-based modeling and argumentation. *Science Scope*, *38*(4), 70–77.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science*, *61*(2), 297–312. http://www.jstor.org/stable/188214

Sandoval, W. (2014). Science education's need for a theory of epistemological development. *Science Education*, *98*(3), 383–387. https://doi.org/10.1002/sce.21107

Shepardson, D. P., Choi, S., Niyogi, D., & Charusombat, U. (2011). Seventh grade students' mental models of the greenhouse effect. *Environmental Education Research*, *17*(1), 1–17. https://doi.org/10.1080/13504620903564549

Sinatra, G. M., & Broughton, S. H. (2011). Bridging reading comprehension and conceptual change in science education: The promise of refutation text. *Reading Research Quarterly*, *46*(4), 374–393. https://doi.org/10.1002/RRQ.005

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist*, *50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924

Sinatra, G. M., Kienhues, D., & Hofer, B. K. (2014). Addressing challenges to public understanding of science: Epistemic cognition, motivated reasoning, and conceptual change. *Educational Psychologist*, *49*(2), 123–138. https://doi.org/10.1080/00461520.2014.916216

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904–1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

State of [MA] Department of Education (2016). *[MA] school performance reports: 2014–2015 school performance reports*. Retrieved from https://xxx.us

[SW] Department of Education (2016). *[SW] annual reports of accountability*. Retrieved from Retrieved from https://xxx.com

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed). Boston, MA: Pearson Education.

Tenenhaus, M., Amato, S., & Esposito Vinzi, V (2004). A global goodness-of-fit index for PLS structural equation modelling. *Proceedings of the XLII SIS Scientific Meeting, Vol. Contributed Papers* (pp. 739–742). Padova, Italy: CLEUP.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y., & Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis*, *48*(1), 159–205. https://doi.org/10.1016/j.csda.2004.03.005

U.S. Environmental Protection Agency. (2017). Wetlands definitions. Retrieved on August 28, 2017 from https://www.epa.gov/cwa-404/section-404-clean-water-act-how-wetlands-are-defined-and-identified

Wang, C. Y. (2015). Scaffolding middle school students' construction of scientific explanations: Comparing a cognitive versus a metacognitive evaluation approach. *International Journal of Science Education*, *37*(2), 237–271. https://doi.org/10.1080/09500693.2014.979378

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, *100*(4), 930–941. https://doi.org/10.1037/a0012842

Wetzels, M., Odekerken-Schroder, G., & van Oppen, C. (2009). Using PLS path modeling for assessing hierarchical construct models: Guidelines and empirical illustration. *MIS Quarterly*, *33*(1), 177–196. Retrieved from https://www.jstor.org/stable/20650284

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3–17. https://doi.org/10.1207/s15326985ep2501_2

How to cite this article: Lombardi D, Bickel ES, Bailey JM, Burrell S. High school students' evaluations, plausibility (re) appraisals, and knowledge about topics in Earth science. *Sci Ed*. 2018;102:153–177. https://doi.org/10.1002/sce.21315

APPENDIX: KNOWLEDGE ITEMS

Topic	Items	Evidence
Climate change	1. The Sun is the main source of energy for Earth's climate.	4
	2. *We cannot know about ancient climate change.*	4
	3. Burning of fossil fuels produces greenhouse gases.	1
	4. Greenhouse gases absorb some of the energy emitted by Earth's surface.	3
	5. Earth's climate is currently changing.	2
Fracking	1. Movement of lithospheric plates causes earthquakes.	3
	2. Fracking cracks dense shale rock formations, which releases natural gas.	4
	3. *It is possible to predict both where and when an earthquake may occur.*	2
	4. Earthquakes are caused by slips in Earth's crustal plates.	3
	5. A rock layer can be porous but impermeable.	1
Wetlands	1. Wetlands occur naturally on every continent.	4
	2. *Loss of wetlands will have little impact on human welfare.*	2
	3. Frogs need wetland habitats in which to reproduce and feed.	1
	4. Draining of some wetlands can result in release of carbon to that atmosphere, which could increase global warming.	3
	5. *Wetlands cause sudden and damaging floods downstream.*	2
Moon	1. Studying Moon rocks brought back by Apollo astronauts can tell us about the history of the Moon.	4
	2. The Moon's orbit around Earth is slightly tilted compared to Earth's orbit around the Sun.	3
	3. *Earth's density is lowest in the core and increases as you go out toward the crust.*	1
	4. Craters on the Moon and mercury show that these objects have been impacted by many smaller objects over billions of years.	2
	5. Scientists have created good models about how the Moon formed.	2

Notes. The final column indicates the evidences (1–4) to which the item corresponds. Italicized items were reverse coded, that is, these are statements with which scientists would disagree.