



Scientific evaluations and plausibility judgements in middle school students' learning about geoscience topics

Timothy G. Klavon, Svetha Mohan, Joshua B. Jaffe, Thalia Stogianos, Donna Governor & Doug Lombardi

To cite this article: Timothy G. Klavon, Svetha Mohan, Joshua B. Jaffe, Thalia Stogianos, Donna Governor & Doug Lombardi (2023): Scientific evaluations and plausibility judgements in middle school students' learning about geoscience topics, *Journal of Geoscience Education*, DOI: [10.1080/10899995.2023.2200877](https://doi.org/10.1080/10899995.2023.2200877)

To link to this article: <https://doi.org/10.1080/10899995.2023.2200877>



[View supplementary material](#) 



Published online: 18 May 2023.



[Submit your article to this journal](#) 



[View related articles](#) 



CrossMark

[View Crossmark data](#) 

Scientific evaluations and plausibility judgements in middle school students' learning about geoscience topics

Timothy G. Klavon^a , Svetha Mohan^b, Joshua B. Jaffe^c, Thalia Stogianos^c, Donna Governor^d and Doug Lombardi^c 

^aCollege of Education and Behavioral Sciences, Black Hills State University, Spearfish, South Dakota, USA; ^bDepartment of Psychology, Tulane University, New Orleans, Louisiana, USA; ^cDepartment of Human Development and Quantitative Methodology, University of Maryland, College Park, Maryland, USA; ^dCollege of Education, University of North Georgia, Dahlonega, Georgia, USA

ABSTRACT

Socially relevant geoscience topics may be difficult for students to learn. For example, connecting hydraulic fracturing to Midwestern US earthquake swarms and using the fossil record to infer past Earth environments may challenge students because of their prior exposures to nonscientific explanations. Sociocognitive theoretical perspectives based on decades of developmental and educational psychology, as well as science education research posit that students may have particular difficulty in evaluating the connections between lines of scientific evidence and explanations. This challenge is especially daunting when students are confronted with various alternative explanations (e.g., scientific and nonscientific explanations). In the present study, we compared two types of scaffolds designed to facilitate Mid-Atlantic middle school students' ($N=40$) scientific thinking and learning about controversial geoscience topics when confronted with alternative explanations. In a less autonomy-supportive scaffold, participants were given four lines of evidence and two explanatory models, one scientific and one nonscientific. (Fracking; *Supplementary Materials 1 & 2*); in a more autonomy-supportive scaffold, students chose four of eight lines of evidence and two of three explanatory models, one scientific and two nonscientific (Fossils; *Supplementary Materials 1 & 2*). Quantitative analyses revealed that both activities facilitated students' evaluations in shifting students' judgments toward the scientific and deepening their knowledge, although the more autonomy-supportive activity had greater effect sizes. Structural equation modeling suggested that more scientific judgments related to greater knowledge at post-instruction for the more autonomy-supportive scaffold. These activities may help students develop more scientific evaluation skills, which are central to understanding geoscience content and science as a process.

ARTICLE HISTORY

Received 25 August 2021

Revised 5 April 2023

Accepted 5 April 2023

KEYWORDS

Scientific thinking;
geoscience knowledge;
instructional scaffolds

Introduction

Reasoning about and evaluating the relations between evidence and alternative explanations about phenomena is fundamental to many scientific practices. In fact, Ford (2015) made the claim that critique and evaluation are foundational to the eight science and engineering practices that are recommended to be incorporated into US K-12 science instruction (e.g., developing and using models; analyzing and interpreting data; NGSS Lead States, 2013). Students may think more critically, behave more scientifically, and gain a deeper understanding of fundamental scientific concepts when engaging in these scientific practices (National Research Council (NRC), 2012). In the present study, we examined the relations between scientific evaluations and judgments about alternative explanatory models, and knowledge of complex and controversial geoscience topics; specifically, (a) the connections between hydraulic fracturing

(fracking) and mid-America earthquake swarms and (b) use of fossil evidence to infer past climatic and surface conditions.

The process of evaluation may be at the core of many, if not all, scientific activities. Recent US science education reform guidance suggested that evaluation—involving both critique and argumentation—is at the nexus between investigating (collecting data and testing solutions about real world phenomena) and developing explanations and solutions (using theories and models to formulate hypotheses and develop solutions) (NRC, 2012). Evaluation is an “iterative process that repeats at every step of [scientific] work” and requires the application of critical thinking skills (NRC, 2012, p. 46). Applied in classroom settings, students may evaluate how scientific data supports alternative hypotheses, and how a specific hypothesis is supported by scientific theories and/or models (Duschl & Bybee, 2014). More

critical evaluations are also a keystone of scientific literacy (Gormally et al., 2012; Jurecki & Wander, 2012, Walsh et al., 2019).

Making scientific evaluations often requires students to reason and be critical about the connections between evidence and explanations. The *Call to Action for Science Education* said that all people should “be able to evaluate evidence and distinguish between what are reliable sources of information, poorly supported claims and unequivocal falsehoods” (National Academies of Science, Engineering and Medicine [NASEM], 2021, p. 15). In the classroom context, students may be scientifically evaluative when asking critical questions, engaging in collaborative arguments, using model-based reasoning, and making more explicit judgments about the plausibility of explanations in light of possible alternatives (Lombardi, Nussbaum et al., 2016). Explicitly reflecting on and forming scientific plausibility judgements may be particularly important in situations when there are competing explanations about a phenomenon (e.g., earthquake swarms in the midwestern US are caused by fossil fuel fracking operations [scientific consensus] vs. natural crustal plate movements [alternative]). Plausibility is specifically an epistemic judgment about the potential truthfulness of an explanation when considering various alternatives (Lombardi, Nussbaum et al., 2016). Lombardi, Bailey et al. (2018) and Lombardi, Bickel et al. (2018) found that when students scientifically evaluate evidence and explanation, and shift their plausibility toward a more scientific stance, they often learn more deeply about geoscience topics (e.g., causes of current climate change and importance of wetlands on ecosystem services). This skill may be linked to students’ critical and argumentative reasoning (Governor et al., 2021; St. John & McNeal, 2017). Further, such skills serve as the foundations for many scientific and engineering practices (Ford, 2015).

Being more scientific and critical in evaluating how well lines of scientific evidence support an explanation—in light of alternatives—is often challenging for scientists and science students alike. In classroom contexts, instructional scaffolding may facilitate more scientific evaluations when evaluating evidence to explanation connections (Kastens & Krumhansl, 2017; Lin et al., 2012; Pea, 2004; Sinatra & Lombardi, 2020). One such scaffold, called the Model-Evidence Link (MEL) diagram, may be particularly effective in facilitating students’ scientific reasoning and evaluations when learning about complex geoscience topics, such as availability of freshwater resources (Medrano et al., 2020). The MEL diagram activities are instructional scaffolds with a diagrammatic structure, designed to facilitate students’ evaluations about the connections between lines of scientific evidence and alternative explanations of a phenomenon (e.g., fracking and fossils; Figure 1; Bailey et al., 2020; Lombardi, 2016). Students then elaborate upon their evaluations in a written explanation task.

The present study used Lombardi, Nussbaum et al.’s (2016) theoretical framework to test the ecological validity and utility (e.g., classroom effectiveness) of MEL scaffolds designed to deepen secondary students’ learning in geoscience. Based on this framework, prior studies have supported the effectiveness of the preconstructed MEL (pcMEL) in

promoting plausibility shifts toward the scientific and reported 5%-10% knowledge gains during the 90-minute lesson (see, for example, Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018). Despite the pcMEL’s effectiveness, students had difficulty transferring their scientific reasoning and knowledge outside the classroom context. Challenges in transfer suggest that students may need to further their agentic engagement in geoscience, where they actively contribute to the flow of instruction and deepen ownership of their learning (LaDue et al., 2022; McNeal et al., 2017; van der Hoeven Kraft, 2017).

Students’ agentic engagement may be facilitated via more autonomy-supportive scaffolding (e.g., instructional scaffolding that supports students’ choice and control). Patall et al. (2019) suggested that agentic engagement may emerge through a dynamic interplay of cognitive, social-behavioral, and emotional factors, which can be sparked via autonomy-supportive practices and scaffolding. Such scaffolding would afford opportunities for students to take action “during a learning experience (begins) by making suggestions, offering input, and expressing preferences” (Reeve & Shin, 2020, p. 152). Thus, our research team developed the build-a-MEL (baMEL) scaffold, which is designed to be more autonomy supportive because students first construct their MEL diagrams by choosing their lines of scientific evidence and alternative explanatory models from sets of options. One previous pilot study suggested that the baMEL was more effective than the pcMEL in promoting students’ scientific evaluations and plausibility judgements and knowledge about water resources (Medrano et al., 2020). The primary goal of the present study was to test the effectiveness of both MEL scaffold types in (a) promoting students’ evaluations when gauging the connections between lines of geoscientific evidence and alternative explanations; (b) promoting plausibility appraisals toward a more scientific stance; and (c) deepening students’ geoscience knowledge about the dynamic nature of Earth’s systems. Based on theoretical and empirical studies in educational and developmental psychology (Lombardi, Nussbaum et al., 2016; Patall et al., 2019; Reeve & Shin, 2020) and science and discipline-based educational research (LaDue et al., 2022), we hypothesized the baMEL activity would show more scientific evaluations and plausibility judgements, as well as deeper knowledge, than the pcMEL.

Literature context

Building upon decades of social cognitive research in science education and developmental and educational psychology, Lombardi, Nussbaum et al. (2016) developed the plausibility judgements in conceptual change model (PJCC). The PJCC posits that plausibility—an epistemic “judgment of potential truthfulness when evaluating explanations (e.g., accounts of phenomena unfold that may lead to a feeling of understanding)—may facilitate learning of complex science topics, particularly when individuals are faced with scientific explanations that conflict with societal normalized, but nonscientific, explanations (e.g., causes of current climate change; Lombardi

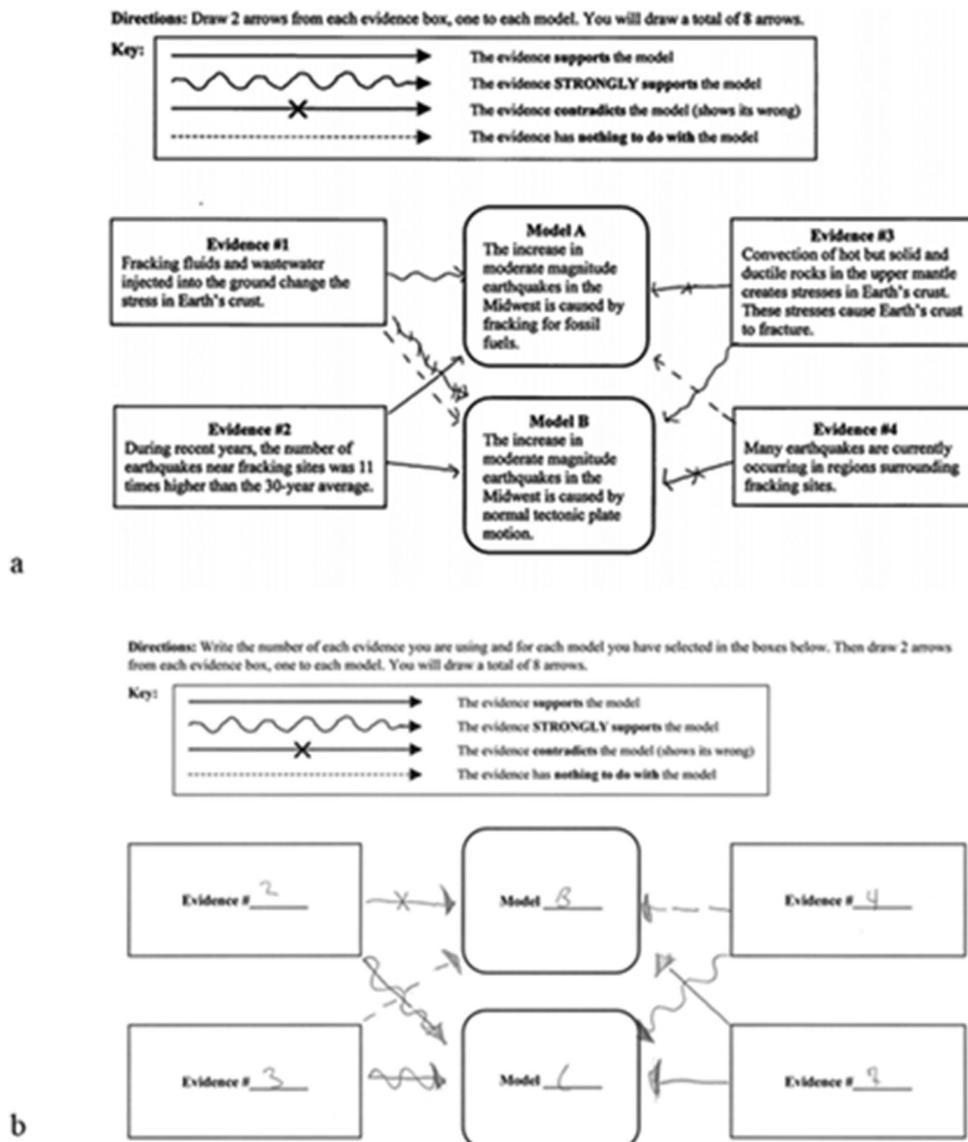


Figure 1. Samples of student work using MEL diagrams.

Note. (a) Student MEL diagram for the Fracking pcMEL activity. (b) Student MEL diagram for the Fossils baMEL activity. The letters denote the models chosen by the student. The numbers denote the lines of evidence chosen by the student.

et al., 2013, p. 35). Operationally, the PJCC model says that students may shift their plausibility toward a more scientific stance and learn more deeply about scientific topics when presented with the opportunity to explicitly evaluate connections between lines of evidence and explanations, in light of alternatives. Continuing research (Bailey et al., 2022; Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018, Medrano et al., 2020) has incorporated the PJCC into the MEL instructional scaffold and supported the PJCC's theoretical position for many Earth science topics, including the climate crisis and availability of freshwater resources.

Scientific evaluations and plausibility judgements

Scientific evaluations are at the core of many scientific and engineering practices. In fact, *A Framework for K-12 Science Education*, on which the *Next Generation Science Standards*

were built (NRC, 2012; NGSS Lead States, 2013), placed “evaluating” at the nexus of the “spheres of activity” for scientists and engineers (p. 45). From a developmental perspective, this framework asserted that—for students to develop scientific literacy—they should make critical scientific evaluations when they reach adolescence. Such critical evaluations often involve interpretations of how well scientific evidence supports competing claims (Ford, 2015). Sinatra and Lombardi (2020) “argued that explicitly reappraising plausibility judgments may be a crucial addition to [critically] evaluating the connections between sources of information [e.g., lines of scientific evidence] and knowledge claims [e.g., scientific explanations of geoscience phenomena]” (p. 128). Empirical evidence suggests that students who express a greater level of evaluation (i.e., from evaluation levels of erroneous to descriptive to relational to critical) when gauging the connections between lines of

scientific evidence and explanatory models, in light of an alternative, are correlational to their shift in plausibility judgments toward a more scientific stance (Lombardi, Brandt, et al., 2016).

Science learning can be facilitated when students engage in scientific evaluations in their classrooms. Chi et al. (2018) posited that greater levels of cognitive engagement occur when science learning is constructive (e.g., when purposefully and critically evaluating the validity of scientific claims). Scientifically critical evaluations involve deeper processing strategies that link and integrate prior knowledge with the alternative and novel science conceptions via elaboration and extension (Arthurs, 2018). However, scientific evaluations between lines of evidence and competing explanations and explicitly appraising and reappraising plausibility judgements requires a relatively high cognitive demand and instructional scaffolding is often needed to help students learn in such situations (Heyd-Metzuyanim & Schwarz, 2017; Lombardi et al., 2021). We specifically designed the MEL scaffolds to support students in making more scientific evaluations and reappraising the plausibility of competing explanatory models to help deepen their learning of complex geoscience concepts and become an active agent of their science learning (Bailey et al., 2020; Lombardi, 2016; Lombardi et al., 2022).

Autonomy to be a science learning agent

Geoscience classrooms are learning communities where students—guided by teachers—make sense of phenomena (e.g., nature and causes of increasing earthquake swarms in the mid-Western US). Early adolescents seek more autonomy in classroom learning communities, particularly when they matriculate through secondary (middle and high) school (Collie, 2020). More autonomy-supportive science classrooms can increase students' agentic engagement as a constructor of science knowledge via participation in scientific discourse and elaborative cognitive processing (Lombardi et al., 2022; Patall et al., 2019). However, such autonomy-supportive environments thrive with properly designed and implemented instructional scaffolding and structure (Lombardi et al., 2021; Reeve, 2013). Recent educational research has revealed that sustained and systematic use of instructional scaffolding can facilitate students reasoning and understanding about complex and controversial scientific phenomena related to Earth's systems and processes (Ceyhan et al., 2019; Darner, 2019; Dauer, 2021; Governor et al., 2021; Medrano et al., 2020; Nussbaum, 2021). For example, previous studies have revealed that the preconstructed form of the MEL scaffold (i.e., the pcMEL) was effective in facilitating students' knowledge construction within the classroom context. This has been suggested by pre- to post-instructional gains and comparative relational pathways derived via structural equation modeling in previous research (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018). However, we wondered if students were more autonomous in constructing their MEL diagrams, would this result in increased agentic engagement. Therefore, our overall purpose was to

investigate the baMEL, which we designed to be a more autonomy-supporting form. We specifically enhanced the MEL scaffold with the hope of increasing students' agentic engagement during the learning process, where students construct their diagrams prior to analyzing and evaluating how well lines of evidence support alternative explanatory models.

The present study

The present study compared two instructional scaffolds: The Fracking pcMEL (Figure 1a; Hopkins et al., 2016) and the Fossils baMEL (Figure 1b; Governor et al., 2020). In the Fracking pcMEL, students are presented with four lines of scientific evidence and two alternative explanatory models about the increased frequency in earthquake activity in the midwestern US in a pre-constructed diagram format. In the Fossils baMEL, students select four lines of scientific evidence from eight possible choices and two alternative explanatory models from three choices about the reliability of subsurface fossils for inferring past paleo-climatic and surface changes. Both activities cover scientific topics that align with several disciplinary core ideas, scientific practices, and crosscutting concepts identified in recent science education reform efforts (NRC, 2012). For example, both scaffolds align with a disciplinary core idea related to Earth's materials and systems (ESS2.A), which says that "Earth's systems, being dynamic and interacting, cause feedback effects that can increase or decrease the original changes" (NRC 2012, p.181). We specifically asked the following research questions:

1. What are the levels of students' evaluations (erroneous, descriptive, relational, or critical) when they engage in the pcMEL and baMEL instructional treatments and how do students' plausibility judgements and knowledge change over the course of these two instructional treatments?
2. How does the increased opportunity for students' agency influence the relations between their levels of evaluation, plausibility judgements, and knowledge when participating in MEL activities?

Based on past research studies and theoretical perspectives on increasing autonomy in classroom learning situations (e.g., Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Patall, 2019), we hypothesized that the MEL scaffolds would result in plausibility shifts toward the scientific and knowledge gains from pre- to post-instruction, with greater levels of evaluation, plausibility shifts, and knowledge gains for the baMEL form of the scaffold (RQ1). Per the PJCC (Lombardi, Nussbaum et al., 2016), we hypothesized that the relational pathway linking levels of evaluation, plausibility, and knowledge would be above and beyond the relational pathway directly linking levels of evaluation and knowledge, as shown in Figure 2. Furthermore, we hypothesized that strength of these relational pathways would be greater with the baMEL based on notions of autonomy-supportive scaffolding (Patall et al., 2019).

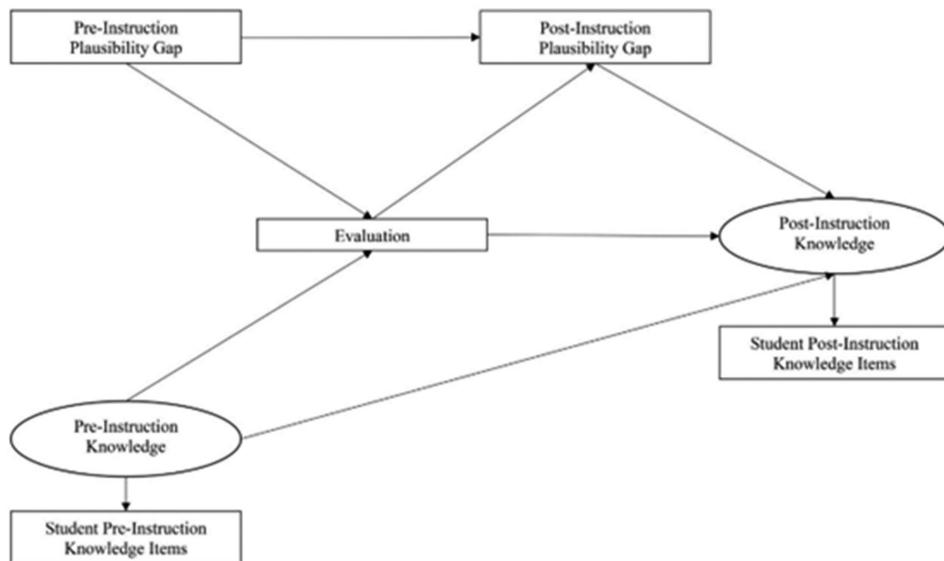


Figure 2. Hypothetical partial least squares-structural equation model relating plausibility, evaluation, and knowledge.

Note. Indicators (i.e., observed values) are designated by rectangles and constructs (i.e., derived values) are designated by ovals. The Fracking pcMEL knowledge score consisted of 5 items. The Fossils baMEL knowledge score consisted of 11 items.

Methods

The present study was situated within a six-year, design-based research project. Such projects typically examine instructional effectiveness through an iterative process (Barab & Squire, 2004). The research team constructed all materials (e.g., lines of evidence and alternative explanations) through a design-based research process, involving iterations of development, testing, analysis, and review. The research team, which includes practicing classroom teachers, science education researchers, educational psychologists, and geoscientists, use primary source documents from highly ranked scientific journals (e.g., *Science*, *Nature*, *Psychological Science in the Public Interest*) in developing these materials. After going through multiple iterations of development and testing, an independent panel of scientific advisors review the materials for accuracy, precision, reliability, and validity. We specifically used pilot test data collected in classroom contexts during the project's second and third years. The data were collected quantitatively or through the quantization of the text of students' written responses. The data were transformed from qualitative to quantitative via the use of a rubric (Lombardi, Brandt et al., 2016) with two coders separately rating each response then coming together to compare rubric scores and reconcile any differences. This process of quantizing facilitates the demonstration of relationships between the evaluation scores and the quantitative variables in the study (Creamer, 2018).

Setting and participants

The present study was exploratory, and we conducted the study at a middle school in the Mid-Atlantic region of the US, with participants ($N=40$) enrolled in a grade 6 Earth

science course of a classroom teacher that attended a previous Earth science education professional development workshop. The participants reflected the demographic characteristics of the school, which is located in a suburban community flanked on one side by a high-density population area of appreciable poverty, and on the other side by a low-density population area of appreciable wealth (US Census, 2021). Specifically, the participants were predominantly Hispanic (of any origin) (41%), with the remainder White (30%), Asian (18%), and Black (11%). Slightly more of the participants were male (52%).

Instructional materials and measures

We implemented the MEL scaffolds, which covered two different geoscience topics: (a) connections between fracking and mid-America earthquake swarms (i.e., Fracking pcMEL) and (b) use of fossil evidence to infer past climatic and surface conditions (i.e., Fossils baMEL). These topics included lines of scientific evidence and explanatory models about geoscience-related phenomena that students could evaluate. Details about the design and structure of each scaffold is found in [Supplementary Materials 1](#) and [2](#). Both topics were part of the curricular scope and sequence in the classroom that participated in the present study.

Fracking pcMEL

The Fracking pcMEL scaffold contained four lines of scientific evidence and two explanatory models (one scientific and one nonscientific alternatives; [Supplementary Materials 1 & 2](#)). Teachers did not tell the students prior to or during the activity which of the two was the scientific model. In the Fracking pcMEL diagram ([Figure 1a](#)), the explanatory models were presented in the center of the page, in two separate

boxes (explanatory texts had no labeling indicating which one is scientific and alternative explanation). The four lines of scientific evidence were located on the left and right edges of the page in boxes. Student participants were also given “evidence texts,” which were one-page summaries for each line of evidence that included expository text, graphs, and/or diagrams reviewed and validated by geoscientists. After reading the texts, the classroom teacher instructed student participants to draw different types of arrows from each evidence text to both models based on how well they thought the evidence supported the model. Four different types of arrows were used; a squiggly arrow indicated the participant believes that the evidence strongly supports the model, a straight arrow indicated that the evidence supports the model, a dotted line arrow indicated the evidence had nothing to do with the model, and a line with an “X” in the middle of it indicated that the evidence contradicts the model. Overall, the participants drew eight arrows in total (Figure 1a).

Fossils baMEL

The Fossils baMEL activity first introduced the students to the eight lines of evidence and the three explanatory models (Supplementary Materials 1 & 2). Unlike the pcMEL diagram, where models are provided, the baMEL diagram template contained two blank boxes in the middle in which participants wrote the letter of the explanatory models they selected (A, B, and/or C), and four blank boxes around the edge where participants wrote the number of the lines of evidence they selected. In other words, participants filled in these blanks by selecting two of the three explanatory models and four of the eight lines of evidence. Compared to the pcMEL, the baMEL gave participants the opportunity to pick which evidence text and explanatory models they would like to incorporate into their MEL diagram and was therefore more autonomy supportive (Reeve & Shin, 2020).

Similar to the pcMEL, students were given one-page “evidence texts” providing more details about each of the eight lines of scientific evidence. After choosing their two models and four lines of evidence, student participants connected the evidence texts to models using the four different types of arrows to connect each evidence and model, for a total of eight total arrows (Figure 1b).

Explanation task: Evaluation score

After completing a MEL diagram (either pcMEL or baMEL), student participants completed what we call the “Explanation Task” (Figure 3). Participants picked two of the connections that they drew from the MEL activity and gave their explanation of why they chose a particular type of arrow to indicate their evaluation of the strength between a particular line of evidence and a particular model. Using a scoring system and rubric developed by Lombardi, Brandt et al. (2016), coders rated explanations for different levels of evaluation using a rubric: 1 = Erroneous, 2 = Descriptive, 3 = Relational, or 4 = Critical (Supplementary Materials 3). These categories represent levels of evaluation based on the accuracy, elaboration, and reasoning present in participants’ responses. To establish coding reliability, two raters independently coded participants’ explanation tasks. They then met and resolved all differences in scoring via discussion, at times with a third coder to assist with that resolution, with full consensus reached after consultation. The final evaluation score was the average of the consensus scores for each explanation.

Model plausibility rating task: Plausibility judgment scores

For both the pcMEL and baMEL, students were instructed to rate the plausibility of all explanatory models, both pre- and post-instruction. For the Fossils baMEL, students

1. Compare more than one evidence text to your models.

Model A is saying the cause which is Fracking is effecting with earth quakes

Model B is saying the cause is natural tectonic plates are causing earth quakes

2. Please work on this part individually after you complete your diagram. Now that you have completed the diagram, reconsider the plausibility of Models A and B (and C, if there is one). Circle the plausibility of each model. [Make one circle for each model.]

		Greatly implausible (or even impossible)									Highly plausible	
		1	2	3	4	5	6	7	8	9	10	
Model A		1	2	3	4	5	6	7	8	9	10	
Model B		1	2	3	4	5	6	7	8	9	10	
Model C (if there is one)		1	2	3	4	5	6	7	8	9	10	

What were your previous ratings? Model A: 5 Model B: 10 Model C (if there is one): _____

3. For the model you selected as most plausible, explain why you think so.

I think Model B is most plausible because the article explained why earthquakes were caused

4. Which arrows changed your plausibility judgments about the models? If your plausibility judgment did not change, which arrows supported your original plausibility judgments? Consider 2 lines of evidence. For each line, does it support, strongly support, or contradict one of the models? Why? When writing your explanation, consider the following:

- Use the specific information from the evidence text and figures to support your response. Ex: when looking at graphs or figures, be sure to describe the patterns in the data
- Describe any cause and effect relationships found in the text.

Evidence # 3 strongly supports | supports | contradicts | has nothing to do with Model B because:

It gave a good understanding of how it happens as well as informs the reader with a visual

Evidence # 4 strongly supports | supports | contradicts | has nothing to do with Model A because:

It explained what fracking does and what effects it has like the earthquakes

5. In your final ranking, did you rank either Model as “1” or “10”? Yes or No [Circle One] Why? Why not?

It was most plausible to me

Figure 3. Sample of student explanation task.
Note. This explanation task is an example from the Fracking pcMEL.

recorded their plausibility judgments for all three explanatory models, while for the Fracking pcMEL, students recorded their plausibility judgments for the two explanatory models. Students gauged the plausibility of each model using a 1–10 scale (1 = greatly implausible and 10 = highly plausible), based on methods used by Lombardi, Sinatra et al. (2013). Because the Fracking pcMEL offered two explanatory models, scores were calculated as the rating of the scientific model minus the alternative. The Fossils baMEL offered three different explanatory models (scientific and two non-scientific alternatives), and therefore, two different scores were calculated: scientific minus nonscientific alternative 1 and scientific minus nonscientific 2. Scores could range on a scale from -9 to +9, where positive scores indicated that participants judged the scientific model as more plausible than the alternative model and negative scores indicated participants judged the nonscientific alternative as being more plausible than the scientific.

Knowledge scores

Student participants completed a multi-item knowledge survey instrument (at pre- and post-instruction). The Fossils Knowledge Survey contained eleven items and the Fracking Knowledge Survey contained five items (Supplementary Materials 4). Students ranked each item on a 5-point Likert scale (1 = strongly disagree and 5 = strongly agree) on their knowledge of how scientists would agree with each item statement per the methods outlined in Lombardi et al. (2013). At least one question in each set addressed each line of scientific evidence. Some questions statements were negatively worded (i.e., in effect scientists would disagree with these knowledge statements) and we reverse coded these statements prior to calculating knowledge scores. McDonald's omega (ω) coefficients were used to examine if the knowledge scale for each scaffold and each time point were sufficiently reliable (Fracking pcMEL pre: $\omega = 0.244$, post: $\omega = 0.723$; Fossils baMEL pre: $\omega = 0.628$, post: $\omega = 0.766$). The Fracking pcMEL may have shown a relatively low omega value at pre-instruction because students have little or no knowledge of fracking prior to the activity, which may indicate that they were guessing about the survey answers randomly (Allen et al., 2008).

Procedures

Figure 4 shows a schematic of the procedures we used. The teacher introduced the Fracking pcMEL activity at the

beginning of an instructional unit. The teacher introduced the Fossils baMEL activity later in the same instructional unit. The order of the activities was determined by instructional needs of the teacher, as well as the providing students activities with increasing levels of complexity. Prior to each activity, students completed a knowledge survey and model plausibility ratings for all explanatory models presented in the activity. At this time, teachers also engaged the class in an unscripted short discussion about the model(s) and the idea of plausibility to clarify misunderstandings and address general questions about the topic. When completing the Fracking pcMEL activity, students read the evidence texts and completed the diagram in small groups. Next, they worked individually to write up the explanation task. The activity ended with the second iteration of the model plausibility ratings and fracking knowledge survey.

When completing the Fossils baMEL activity, students first read the texts for all eight lines of evidence and then were introduced to the three alternative models explaining the phenomenon. Small groups of students worked together to select four lines of evidence from the eight available and two alternative models from the three available. The students used the four lines of evidence and two models to construct a MEL diagram, which they then completed by drawing arrows for both activities. In both the Fracking pcMEL and the Fossils baMEL activities, students worked individually to complete the explanation task. Each activity ended with the second iteration of the model plausibility ratings and corresponding knowledge survey. Upon completion of this sequence, teachers moved on to teaching their regular instructional unit. Each MEL activity took place over about two regular class periods (~90 minutes total), with negligible difference in instructional time between the Fracking pcMEL and Fossils baMEL.

Results

We present the results in two sections—the first addresses RQ1: How do students' (a) evaluations about the connections between lines of scientific evidence and explanatory models, (b) shifts in plausibility judgements, and (c) changes in knowledge about geology compare between the two instructional scaffolds, the Fracking pcMEL and the Fossils baMEL?; and the second addresses RQ2: How do the relationships between plausibility, evaluation, and knowledge compare between the Fracking pcMEL and the Fossils baMEL? The analyses associated with RQ1 were paired-samples comparisons between the instructional treatments, while those

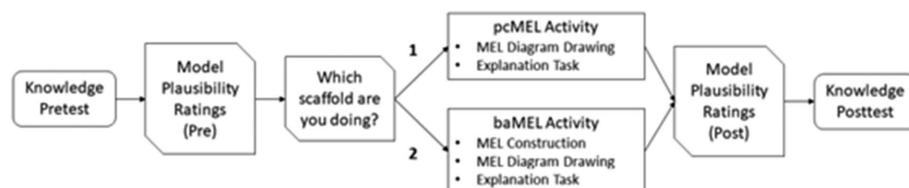


Figure 4. Student procedures for completing a MEL diagram.

Note. Students first completed the knowledge survey and the plausibility ratings. Then, students completed each pcMEL (1) and baMEL (2) activity over about two regular class periods, after which they completed another knowledge survey and plausibility ratings. The only difference between the two MEL activities is that in the baMEL, students chose the models and evidence statements from a set of options.

associated with RQ2 were structural relations modeled between the three variables: levels of evaluation, plausibility judgements, knowledge. Prior to conducting these analyses, we screened the data for outliers and to ascertain alignment with normality assumptions, which is common in many quantitative educational research studies. We found no univariate outliers and concluded sample normality after examining skewness, kurtosis, and normal probability plots.

Research Question 1

Although the screening analysis showed that the sample was reasonably normal, we conducted a categorical nonparametric data analysis because we knew ahead of time the sample size was relatively small. Therefore, we did not assume—a priori—that the sampling distribution would approach normality because Nussbaum (2014) suggested that choosing a statistical test (e.g., a *t*-test, which is commonly used for relatively large sample sizes) prior to data analyses is scientifically appropriate. Another reason we chose to conduct a categorical nonparametric test was to optimize the potential power of the analysis on this relatively small sample. We specifically conducted Wilcoxon signed-ranks tests, to investigate differences in levels of evaluation and pre- to post-instructional differences in plausibility and knowledge scores between the Fracking pcMEL and the Fossils baMEL. Shown in more detail in Table 1 and Figures 5–7, the Wilcoxon signed-ranks tests revealed that evaluation scores were significantly greater for the Fossils baMEL than for the Fracking pcMEL, with medium effect size (Kerby, 2014). There was also a significant shift in plausibility scores toward the scientific for the pcMEL, with a large effect size. However, there was not a significant shift in plausibility toward the scientific model for the baMEL. There were significant increases in knowledge scores for both the pcMEL and baMEL, with both gains showing a medium effect size.

Research Question 2

We used structural equation modeling to answer Research Question 2. Prior to conducting the analyses, we constructed a hypothetical conceptual model based on Lombardi, Nussbaum et al.'s (2016) theoretical framework linking evaluation, plausibility, and knowledge, as well as prior empirical research (Figure 2; see for example, Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018). The conceptual model reflects this research question that we want to test empirically. The conceptual model also could be directly adopted as our measurement model in a structural equation

modeling (SEM) environment. Figure 2 specifically depicts a model where plausibility judgments mediate the relations between levels of evaluation and knowledge. The model also reflects the direct relation between levels of evaluation and knowledge. Therefore, in the subsequent analyses, we were able to compare these two different pathways.

We first analyzed these relations using WarpPLS 7.0 (Kock, 2020), a program that employs a “warping” partial least squares (PLS) structural equation modeling (SEM) path analysis to afford greater accuracy by not assuming linear relations between variables (Kock, 2016).

We selected jackknifing as the resampling technique when running our path analyses because this technique may reduce standard error and increase statistical power for relatively small sample sizes by removing one or more indicators at a time and replacing them with partial estimates (Abdi & Williams, 2010; Quenouille, 1949; Tukey, 1958). Using this jackknifing replacement method often increases the predictive ability of the PLS-SEM (Kock, 2020). Model comparisons were made using Tenenhaus Goodness of Fit (GoF), which answers how well different subsets of the data can be explained by the model and is an indicator of overall model robustness and strength (Henseler & Sarstedt, 2013). WarpPLS constructed the students' pre- and post-instruction knowledge scores latent variables using the individual knowledge survey items as indicators.

After completing the PLS-SEM, we implemented a holistic approach to evaluating the relationships formed by the model, using the significance, the standardized path values, and the effect size of each pathway (measured by Cohen's *f*-squared; Smith, 2020). Though significance testing via *p*-value played an important role in how we assessed our data, recent guidance suggests that *p* values alone should not exclude relationships, in the light of additional pathway strength indices (Amrhein et al., 2019). Further, Wasserstein et al. (2019) implored researchers to not “believe that an association or effect is absent just because it was not statistically significant” (p. 1). Therefore, this more holistic approach provided us the opportunity to more fully gauge the strength and robustness of pathway relations than is afforded by using *p*-value alone.

Fracking pcMEL PLS-SEM results

We found that the Fracking pcMEL produced a robust and strong fit (Tenenhaus GoF = 0.453, large effect size). Standardized path values and pathway effect sizes revealed strong relations from pre-instruction plausibility (PrP) to evaluation (E) and to post-instruction plausibility (PoP). The E to PoP relationship is strong, as is the pre-instruction

Table 1. Pre- to post-instructional variable score differences.

	pc-MEL		ba-MEL		<i>p</i>	<i>r</i>
	<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)	<i>T</i>	<i>p</i>		
Evaluation	2.03 (0.70)	2.29 (0.76)	256	.038	.462	
Plausibility	pc- MEL pre <i>M</i> (<i>SD</i>)	ba- MEL post <i>M</i> (<i>SD</i>)	<i>T</i>	<i>p</i>	<i>r</i>	pc-MEL <i>p</i>
Knowledge	-2.58 (4.23)	2.68 (3.48)	693	< .001	.970	pre <i>M</i> (<i>SD</i>)
	3.58 (0.51)	3.79 (0.43)	403	.028	.437	post <i>M</i> (<i>SD</i>)
						<i>T</i>
						<i>p</i>
						<i>r</i>

Note. *M* = score mean, *SD* = score standard deviation, *T* = the test statistic for the Wilcoxon signed-ranks test (the sum of the signed ranks), and *r* = matched ranked biserial correlation coefficient, which is a measure of effect size for the Wilcoxon signed-ranks test.

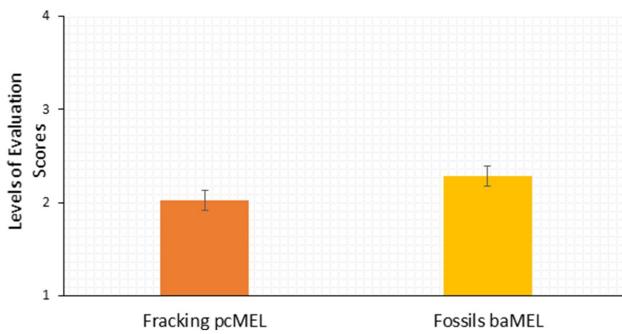


Figure 5. Levels of evaluation scores for the two scaffolds.
Note. $N=40$. Error bars indicate ± 1 SE.

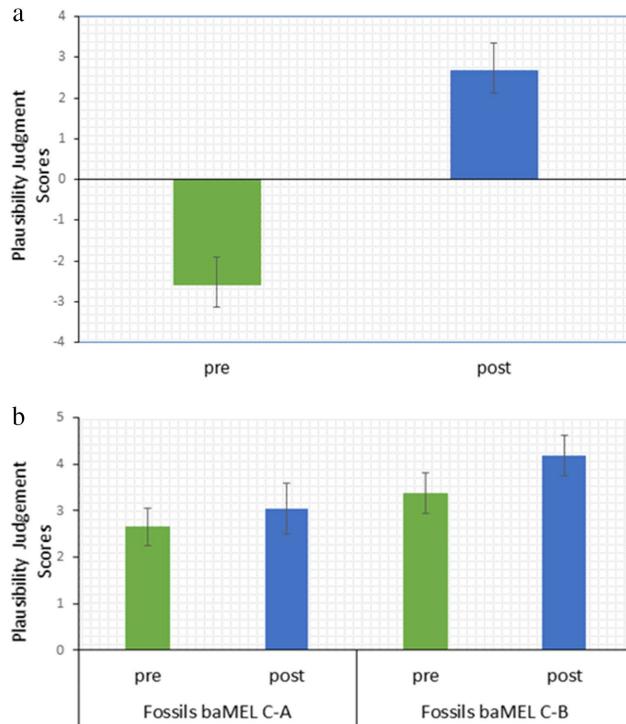


Figure 6. Pre- to post-instruction plausibility judgment scores for the two scaffolds.
Note. $N=40$. (a) Fracking pcMEL and (b) Fossils baMEL. Error bars indicate ± 1 SE.

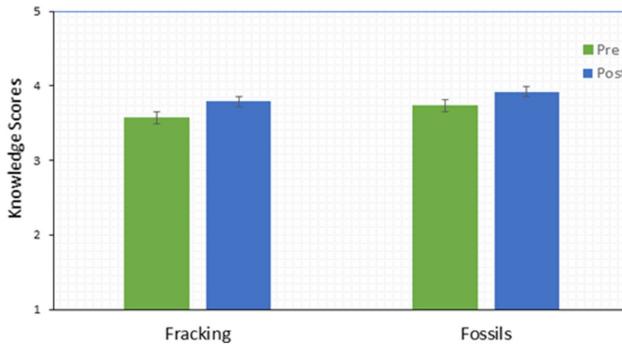


Figure 7. Pre- to post-instruction knowledge scores for the two scaffolds.
Note. $N=40$. Error bars indicate ± 1 SE.

knowledge (PrK) to post-instruction knowledge (PoK) relationship (Table 2). The remaining links between variables were relatively weak (Figure 8).

Table 2. PLS-structural equation modeling β weights, effect sizes, and significance values for the fracking pcMEL relationships.

	Pre-Instruction Plausibility			Post-Instruction Plausibility			Pre-Instruction Knowledge		
	β	f	p	β	f	p	β	f	p
Evaluation	0.485	0.239	.018	—	—	—	—	—	0.005
Post-Instruction	0.414	0.212	.000	0.314	0.139	.025	—	—	—
Plausibility	—	—	—	—	—	—	—	—	—
Post-Instruction	—	—	—	0.191	0.021	.316	0.193	0.031	.309
Knowledge	—	—	—	—	—	—	0.478	0.209	.020

Note. $N=40$. β represents standardized pathway weights, f^2 represents the WarpPLS 7.0 approximation of Cohen's f^2 as an indicator of effect size, and p represents p -value.

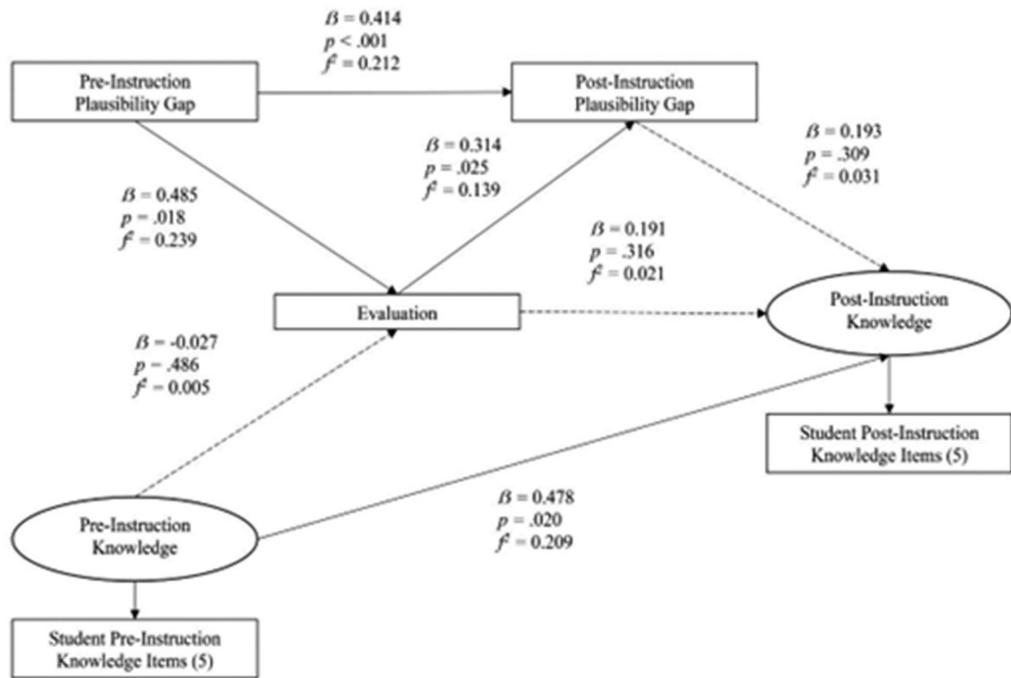


Figure 8. Partial least squares-structural equation model relating the fracking pcMEL plausibility, evaluation, and knowledge.

Note. $N=40$. Indicators (i.e., observed values) are designated by rectangles and constructs (i.e., derived values) are designated by ovals. Solid lines denote strong relationships and dashed lines denote weak relationships. The Fracking pcMEL knowledge score consisted of 5 items.

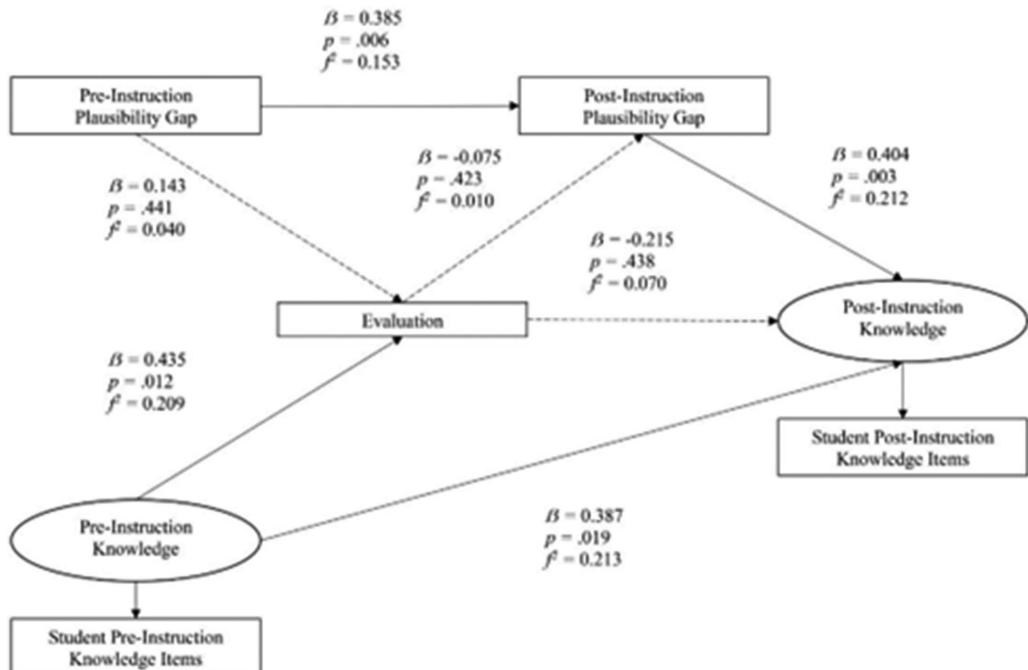


Figure 9. Partial least squares-structural equation model relating the fossils baMEL plausibility, evaluation, and knowledge.

Note. $N=40$. Indicators (i.e., observed values) are designated by rectangles and constructs (i.e., derived values) are designated by ovals. Solid lines denote strong relationships and dashed lines denote weak relationships. The Fossils knowledge score consisted of 11 items.

The strength of the PrK-PoK relationship is not surprising. Past research (see, for example, Braasch & Goldman, 2010; Klosterman & Sadler, 2010) has indicated strong relations between students' prior knowledge and their post-instruction knowledge gains. This is particularly evident when analogies (which are often used in explanatory models) provide the framework for that construction because

these analogies provide a frame of reference for the development of mental models (Norman, 1983). The E-PoP relationship is also strong, though not quite as strong as either the PrP-E or the PrP-PoP relationships. The E-PoP relationship is not surprising, as previous research has indicated that the evaluation of the relations between evidence and explanatory models does impact students' plausibility

reappraisals (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Medrano et al., 2020). The PoP-PoK relation was not robust, which contrasts with previous empirical work investigating the pcMEL for other topics (Medrano et al., 2020).

Fossils baMEL PLS-SEM results

The Fossils baMEL PLS-SEM (Figure 9) also exhibited a robust and strong goodness of fit (Tenenhaus GoF = 0.427, large effect size). The PrK-PoK and PrK-E relations were of moderate strength, as was the PrP-PoP relationship (Table 3). The PoP-PoK relationship is of note, due to the relatively high standardized path value ($\beta = 0.40$) and moderate effect size ($f^2 = 0.21$). Students often have extensive preexisting knowledge about fossils, even from an early age (Borgerding & Raven, 2018), and this may lead to the impact of prior knowledge on the PoK score.

PLS-SEM comparison

We use two metrics to compare the two PLS-SEM outcomes: GoF, as an indicator of overall SEM effect size, and Cohen's f -squared, as indicators of specific pathway effect sizes. Both the Fracking pcMEL and the Fossils baMEL exhibit strong effect sizes (Tenenhaus GoF = 0.453 and 0.427, respectively) that are virtually equivalent and suggest that the overall model structures are very similar. As shown in Table 4, both PLS-SEMs show a relatively strong PrK-PoK relationship with moderate effect size, as well as PrP-PoP relationships of similar importance. The Fracking pcMEL does exhibit a stronger E-PoP relationship than the Fossils baMEL, whereas the baMEL has a much stronger PoP-PoK relationship. Figure 10 shows these comparisons graphically, with a 1:1 comparison shown as a diagonal line on the graph as a reference. Points above and to the left of this reference reveal a pathway comparison that favors the baMEL. Points below and to the right of the reference reveal a pathway comparison that favors the pcMEL. The PoP-PoK relationship comparison shows that students' post-instructional plausibility judgements have a greater impact on post-instructional knowledge gains for the baMEL. Conversely, the PrK-E relationship is more dominant with the pcMEL. This suggests that the baMEL is more favorable in promoting plausibility shifts, with students relying less on their background knowledge—which may be inconsistent with geoscientific understanding—with the pcMEL. However, both forms of the scaffold showed knowledge increases, which indicates that the relations between levels of evaluation, plausibility shifts, and knowledge gains may be somewhat more complex than the hypothesized model.

Discussion

Although somewhat mixed, the results suggest that the baMEL activity is more effective than the pcMEL. We hypothesized that the baMEL activity would show more scientific evaluations and plausibility judgements, and

	Pre-Instruction Plausibility						Post-Instruction Plausibility						Post-Instruction Knowledge					
	β	f^2	p	β	f^2	p	β	f^2	p	β	f^2	p	β	f^2	p			
Evaluation	0.143	0.040	.441	—	—	—	—	—	—	—	—	—	0.435	0.209	.012			
Post-Instruction	0.385	0.153	.006	—0.075	0.010	.423	—	—	—	—	—	—	—	—	—			
Plausibility	—	—	—	—0.215	0.070	.438	0.404	0.212	0.003	0.387	0.123	.019	—	—	—			
Post-Instruction	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—			
Knowledge	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—			

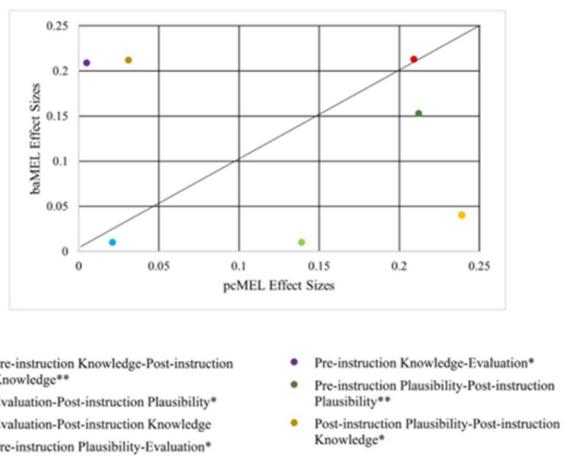
Table 3. PLS-structural equation modeling β weights, effect sizes, and significance values for the fossils baMEL relationships.

Note. $N = 40$. β represents standardized pathway weights, f^2 represents the WarpPLS 7.0 approximation of Cohen's f^2 as an indicator of effect size, and p represents p -value.

Table 4. Effect size comparison.

Link Description	Effect Size (F^2)	
	pcMEL	baMEL
Pre-instruction Plausibility-Evaluation	0.239*	0.040
Pre-instruction Plausibility-Post-instruction Plausibility	0.212*	0.153*
Pre-instruction Knowledge-Evaluation	0.005	0.209*
Pre-Instruction Knowledge-Post-instruction Knowledge	0.209*	0.213*
Evaluation-Post-instruction Plausibility	0.139*	0.010
Post-instruction Plausibility-Post-instruction knowledge	0.031	0.212*
Evaluation-Post-instruction Knowledge	0.021	0.010

Note. $N=40$. *Link included in final PLS-SEM.

**Figure 10.** Comparison of approximate Cohen's f-squared effect sizes.

Note. $N=40$. *Denotes link included in one PLS-SEM. **Denotes link included in both PLS-SEMs.

deeper knowledge than the pcMEL. The analyses associated with RQ1 revealed that the baMEL resulted in moderately greater levels of evaluation than the pcMEL, with both scaffolds moderately increasing students' geoscience knowledge from pre- to post-instruction. In terms of practical significance, students increased their knowledge for each topic by about 5%, which is meaningful and practically significant for classroom use (Figures 4–6). Given that the MEL activities take only about 90 minutes and constitute just one in a series of lessons that instructors would probably use in a geoscience instructional unit, they may represent an effective use of classroom instruction time. However, only the pcMEL resulted in moderate shifts toward more scientific plausibility judgments, pre- to post-instruction. This result may suggest that students have different levels of prior knowledge about fracking/earthquakes and fossils. Based on the learning progression found in *A Framework for Science Education*, upon which the NGSS and many state science education standards have been constructed, students may have had little or no instruction about fracking/earthquakes prior to middle school, whereas students may have had appreciable instruction about fossils in elementary school (NRC, 2012). The framework also suggests that fossils and paleoclimatology

are important concepts to be learned in elementary school, with increased emphasis in middle school and beyond (Borgerding & Raven, 2018; Governor et al., 2020). For example, fossils are specifically referenced in multiple elementary disciplinary core ideas within the standards. However, fracking is not explicitly mentioned at any level. Although teachers may cover certain aspects of natural resources, such as fracking and fossil fuel drilling, instruction related to fracking is minimally supported by the NGSS and state standards frameworks prior to middle school (NGSS Lead States, 2013). Thus, the novelty of the topic (in this case fracking/earthquakes) may be a factor in influencing plausibility appraisal and (re) appraisal (Governor et al., 2020). Another plausible explanation may be that, because the pcMEL was introduced to the participants first, the novelty of engaging in the scaffold may have influenced their depth of scientific reasoning or influenced the outcome of the baMEL activity (Lombardi et al., 2022). Either of these explanations may be supported when considering the participants' age was at the beginning of early adolescence, a time of transition in science learning from more concrete to more abstract conceptual representations (Driver & Easley, 1978). In past empirical studies, high school students showed strong plausibility judgment shifts, pre- to post-instruction, when using the Fracking pcMEL (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018).

Results from the structural pathways comparison (RQ2) further support our hypothesis. Both the Fracking pcMEL and the Fossils baMEL PLS-SEMs suggested that post-instructional knowledge may have been driven by the students' prior knowledge and, to a lesser extent, their pre-instruction plausibility judgements. This is not overly surprising, as previous research has long shown that prior knowledge is quite influential in students' learning contexts (Alexander et al., 1994; McCarthy & McNamara, 2021). When students possess greater prior knowledge about a subject, they are more likely to have frameworks that allow for greater engagement with, and deeper understanding of activities designed for additional learning within that subject (Bae & DeBusk-Lane, 2019; Sinatra et al., 2015). Therefore, there is a stronger foundation upon which new knowledge may be built. However, these comparisons also indicated that students' post-instructional plausibility judgements for the Fossils baMEL showed a much stronger relation to post-instructional knowledge, compared to the Fracking pcMEL. Although results from the paired sample tests did not show meaningful shifts in plausibility pre- to post-instruction, those students that had greater levels of evaluation expressed more scientific plausibility judgements and deeper knowledge at post-instruction with the Fossils baMEL. In light of past research (Bailey et al., 2022; Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Medrano et al., 2020), these greater levels of evaluation may be supporting plausibility shifts not necessarily evident in this small sample. This could provide some support for Patall et al.'s (2019) and Reeve and Shin's (2020) ideas about the effectiveness of autonomy-supportive scaffolding. All in all, the performance of the Fossils baMEL is encouraging as our research team's efforts progress in developing and

testing instructional scaffolding to effectively facilitate geoscience and environmental science instruction.

Limitations

K-12 classrooms are complex learning environments and conducting ecologically valid studies in such settings is often challenging. We acknowledge these constraints, but also try to contextualize the present study within the larger research program represented by our team's efforts and supported by the US National Science Foundation (e.g., Bailey et al., 2022; Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Medrano et al., 2020). For example, this study's participant sample—which was predominantly Hispanic (of any origin)—was conducted in one classroom, at one school. Therefore, some caution is warranted in trying to generalize these results beyond this particular context. Some consideration also needs to be given to the MEL topics. For example, on the one hand, the topic of fossils is found in many places in the NGSS (NGSS Lead States, 2013), with fossils, a common phenomenon, found throughout the world. On the other hand, the topic of earthquakes is mentioned in the NGSS relatively less frequently, with fracking operations being a more regional phenomenon (e.g., the midwestern US) and not mentioned in the NGSS at all. Further, although we are ostensibly comparing the effectiveness of two iterations of the MEL diagram activity, these two geoscience MELs cover separate topics. As the MEL diagrams themselves were completed in small groups, there may have been social impacts upon individual student's learning. This phenomenon has been documented in previous studies (Governor et al., 2021; Lombardi et al., 2022) and may strengthen the knowledge gains among the participants. Our inclusion of pre- and post-instructional measures in both the paired sample tests largely control for this topic difference, but we do acknowledge that there may be some confounding effects in this regard, albeit likely quite minor. Given that this was a pilot study, we plan to incorporate our findings and limitations into a more rigorous quasi-experimental design that will further limit any confounding factors.

Implications and conclusion

The results of the present study revealed that instructional scaffolding can facilitate students to have (a) deeper levels of evaluation between lines of evidence and alternative explanatory models, (b) plausibility shifts toward the scientific alternative, and (c) increased understanding of geoscience. Although both types of scaffolds were effective in this regard, the Fossils baMEL, which is more autonomy-supportive, may have been slightly advantageous because it resulted in strong relations between levels of evaluation and post-instructional plausibility judgments and knowledge. Continued research using these scaffolds will help to better identify the relationships between these three constructs to better define the role of plausibility shifts in the baMEL activities. Implementing these tools in a classroom context

may require a sequential approach (i.e., introducing a pcMEL first), so that students can deepen their understanding of the process of scientific evaluation and judgment making, and then introducing baMELs, which might afford the students more agency in their scientific knowledge construction.

Additionally, strategies may need to be developed and explored to improve the overall efficacy of the MEL activities to maximize their potential in developing critical reasoning skills for evaluating evidence-based claims about scientific phenomena. For example, the process of debating claims and evidence with peers is a critical part of the scientific practice of constructing knowledge (NASEM, 2021). However, engaging in constructive scientific discourse requires the use of negotiation practices that educators may need to model and facilitate (Governor et al., 2021), which might enhance the use of all instructional strategies designed for students to collaboratively evaluate claims and models. Overall, providing opportunities for evaluating claims and evidence through the MEL instructional scaffolds may help students understand the how and why of science, along with core scientific principles—knowledge that in turn, could deepen their scientific literacy and position them to be more effective problem solvers and agents of change in their local, regional, and global communities.

Acknowledgments

The authors would like to acknowledge Janelle M. Bailey for her contribution to the Methods section of this manuscript.

Funding

This research was supported, in part, by the U.S. National Science Foundation (NSF) under Grant No. 2027376. Any opinions, findings, conclusions, or recommendations expressed are those of the author and do not necessarily reflect the NSF's views.

ORCID

Timothy G. Klavon  <http://orcid.org/0000-0002-2890-0970>
 Doug Lombardi  <http://orcid.org/0000-0002-4172-318X>

References

- Abdi, H., & Williams, L. J. (2010). Jackknife. In N. Salkind (Ed.), *Encyclopedia of research design*. Sage.
- Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). How subject-matter knowledge affects recall and interest. *American Educational Research Journal*, 31(2), 313–337. <https://doi.org/10.3102/00028312031002313>
- Allen, K., Reed-Rhoads, T., Terry, R. A., Murphy, T. J., & Stone, A. D. (2008). Coefficient alpha: An engineer's interpretation of test reliability. *Journal of Engineering Education*, 97(1), 87–94. <https://doi.org/10.1002/j.2168-9830.2008.tb00956.x>
- Amrhein, V., Greenland, S., & McShane, B. (2019, March 20). Scientists rise up against statistical significance. *Nature*, 567(7748), 305–307. https://www.nature.com/articles/d41586-019-00857-9?fbclid=IwAR258v566K49c_cig0f9FhPov_UnuGft3DZ0YbM095emw4iK87Zmz1ovY0s <https://doi.org/10.1038/d41586-019-00857-9>

Arthurs, L. (2018). How explicit is the cognitive science foundation of geoscience education research? A study of syntactical units in JGE articles. *Journal of Geoscience Education*, 66(1), 77–91. <https://doi.org/10.1080/10899995.2018.1411710>

Bae, C. L., & DeBusk-Lane, M. (2019). Middle school engagement profiles: Implications for motivation and achievement in science. *Learning and Individual Differences*, 74, 101753. <https://doi.org/10.1016/j.lindif.2019.101753>

Bailey, J. M., Jamani, S., Klavon, T. G., Jaffe, J., & Mohan, S. (2022). Climate crisis learning through scaffolded instructional tools. *Educational and Developmental Psychologist*, 39(1), 85–99. <https://doi.org/10.1080/20590776.2021.1997065>

Bailey, J. M., Klavon, T. G., & Dobaria, A. (2020). The Origins build-a-MEL: Introducing a scaffold to explore the origins of the Universe. *The Earth Scientist*, 36(3), 7–11.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, 13(1), 1–14. <https://doi.org/10.4324/9780203764565-0>

Borgerding, L. A., & Raven, S. (2018). Children's ideas about fossils and foundational concepts related to fossils. *Science Education*, 102(2), 414–439. <https://doi.org/10.1002/sce.21331>

Braasch, J. L., & Goldman, S. R. (2010). The role of prior knowledge in learning from analogies in science texts. *Discourse Processes*, 47(6), 447–479. <https://doi.org/10.1080/01638530903420960>

Ceyhan, G. D., Mugalolu, E. Z., & Tillotson, J. W. (2019). Teaching socio-scientific issues through evidence-based thinking practices: Appropriateness, benefits, and challenges of using an instructional scaffold. *İlköğretim Online*, 18(4), 1405–1417. <https://doi.org/10.17051/ilkonline.2019.630305>

Chi, M. T. H., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., Levy, R., Li, N., McEldoon, K. L., Stump, G. S., Wylie, R., Xu, D., & Yaghmourian, D. L. (2018). Translating the ICAP theory of cognitive engagement into practice. *Cognitive Science*, 42(6), 1777–1832. <https://doi.org/10.1111/cogs.12626>

Collie, R. J. (2020). The development of social and emotional competence at school: An integrated model. *International Journal of Behavioral Development*, 44(1), 76–87. <https://doi.org/10.1177/0165025419851864>

Creamer, E. G. (2018). *An introduction to fully integrated mixed methods research*. SAGE Publications.

Darner, R. (2019). How can educators confront science denial? *Educational Researcher*, 48(4), 229–238. <https://doi.org/10.3102/0013189X19849415>

Dauer, J. M., Sorensen, A. E., & Wilson, J. (2021, May) Students' civic engagement self-efficacy varies across socioscientific issues contexts. *Frontiers in Education*, 6, 154. <https://doi.org/10.3389/feduc.2021.628784>

Driver, R., & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students. *Studies in Science Education*, 5(1), 61–84. <https://doi.org/10.1080/03057267808559857>

Duschl, R. A., & Bybee, R. W. (2014). Planning and carrying out investigations: An entry to learning and to teacher professional development around NGSS science and engineering practices. *International Journal of STEM Education*, 1(1), 1–9. <https://doi.org/10.1186/s40594-014-0012-6>

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>

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE Life Sciences Education*, 11(4), 364–377. <https://doi.org/10.1187/cbe.12-03-0026>

Governor, D., Lombardi, D., & Duffield, C. (2021). Negotiations in scientific argumentation: An interpersonal analysis. *Journal of Research in Science Teaching*, 58(9), 1389–1424. <https://doi.org/10.1002/tea.21713>

Governor, D., Strickland, K., & Bailey, J. M. (2020). Climate changes of the past: Engaging in evidence-based argumentation. *The Earth Scientist*, 36(3), 13–17.

Henseler, J., & Sarstedt, M. (2013). Goodness-of-fit indices for partial least squares path modeling. *Computational Statistics*, 28(2), 565–580. <https://doi.org/10.1007/s00180-012-0317-1>

Heyd-Metzuyanim, E., & Schwarz, B. B. (2017). Conceptual change within dyadic interactions: The dance of conceptual and material agency. *Instructional Science*, 45(5), 645–677. <https://doi.org/10.1007/s11251-017-9419-z>

Hopkins, J., Crones, P., Burrell, S., Bailey, J. M., & Lombardi, D. (2016). Evaluating the connections between fracking and earthquakes. *The Earth Scientist*, 32(2), 23–30.

Jurecki, K., & Wander, M. C. (2012). Science literacy, critical thinking, and scientific literature: Guidelines for evaluating scientific literature in the classroom. *Journal of Geoscience Education*, 60(2), 100–105. <https://doi.org/10.5408/11-221.1>

Kastens, K., & Krumhansl, R. (2017). Identifying curriculum design patterns as a strategy for focusing geoscience education research: A proof of concept based on teaching and learning with geoscience data. *Journal of Geoscience Education*, 65(4), 373–392. <https://doi.org/10.5408/16-217.1>

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3, 11-IT. <https://doi.org/10.2466/11.IT.3.1>

Klosterman, M. L., & Sadler, T. D. (2010). Multi-level assessment of scientific content knowledge gains associated with socioscientific issues-based instruction. *International Journal of Science Education*, 32(8), 1017–1043. <https://doi.org/10.1080/09500690902894512>

Kock, N. (2016). Non-normality propagation among latent variables and indicators in PLS-SEM simulations. *Journal of Modern Applied Statistical Methods*, 15(1), 299–315. <https://doi.org/10.22237/jmasm/1462076100>

Kock, N. (2020). *WarpPLS user manual: Version 7.0*. ScriptWarp Systems.

LaDue, N. D., McNeal, P. M., Ryker, K., St. John, K., & van der Hoeven Kraft, K. J. (2022). Using an engagement lens to model active learning in the geosciences. *Journal of Geoscience Education*, 70(2), 144–160. <https://doi.org/10.1080/10899995.2021.1913715>

Lin, T. C., Hsu, Y. S., Lin, S. S., Changlai, M. L., Yang, K. Y., & Lai, T. L. (2012). A review of empirical evidence on scaffolding for science education. *International Journal of Science and Mathematics Education*, 10(2), 437–455. <https://doi.org/10.1007/s10763-011-9322-z>

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., Bailey, J. M., Bickel, E. S., & Burrell, S. (2018). Scaffolding scientific thinking: Students' evaluations and judgments during Earth science knowledge construction. *Contemporary Educational Psychology*, 54, 184–198. <https://doi.org/10.1016/j.cedpsych.2018.06.008>

Lombardi, D., Bickel, E. S., Bailey, J. M., & Burrell, S. (2018). High school students' evaluations, plausibility (re) appraisals, and knowledge about topics in Earth science. *Science Education*, 102(1), 153–177. <https://doi.org/10.1002/sce.21315>

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., Matewos, M. M., Jaffe, J., Zohery, V., Mohan, S., Bock, K., & Jamani, S. (2022). Discourse and agency during scaffolded middle school science instruction. Online publication. *Advance*, 59(5–6), 379–400. *Discourse Processes*. <https://doi.org/10.1080/0163853X.2022.2068317>

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., Shipley, T. F., Bailey, J. M., Bretones, P. S., Prather, E. E., Ballen, C. J., Knight, J. K., Smith, M. K., Stowe, R. L., Cooper, M. M., Prince, M., Atit, K., Uttal, D. H., LaDue, N. D., McNeal, P. M., Ryker, K., St. John, K., van der Hoeven Kraft, K. J., & Docktor, J. L. (2021). The curious construct of active learning. *Psychological Science in the Public Interest*, 22(1), 8–43. <https://doi.org/10.1177/1529100620973974>

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>

McCarthy, K. S., & McNamara, D. S. (2021). The multidimensional knowledge in text comprehension framework. *Educational Psychologist*, 56(3), 196–214. <https://doi.org/10.1080/00461520.2021.1872379>

McNeal, K. S., John, K. S., Kortz, K., Nagy-Shadman, E., & Riggs, E. (2017). Introduction to the theme: Synthesizing results and defining future directions of geoscience education research. *Journal of Geoscience Education*, 65(4), 347–352. <https://doi.org/10.5408/1089-9995-65.4.347>

Medrano, J., Jaffe, J., Lombardi, D., Holzer, M. A., & Roemmele, C. (2020). Students' scientific evaluations of water resources. *Water*, 12(7), 2048. <https://doi.org/10.3390/w12072048>

National Academies of Sciences, Engineering and Medicine (NASEM). (2021). *Call to action for science education: Building opportunity for the future*. The National Academies Press. <https://doi.org/10.17226/26152>

National Research Council (NRC). (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. <https://doi.org/10.17226/13165>

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press. <http://www.nextgenscience.org/>

Norman, D. A. (1983). Some observations on mental models. In D. Gentner and A. L. Stevens (Eds.), *Mental models* (pp. 7–14). Lawrence Erlbaum Associates.

Nussbaum, E. M. (2014). *Categorical and nonparametric data analysis: Choosing the best statistical technique*. Routledge.

Nussbaum, E. M. (2021). Critical integrative argumentation: Toward complexity in students' thinking. *Educational Psychologist*, 56(1), 1–17. <https://doi.org/10.1080/00461520.2020.1845173>

Patall, E. A., Pituch, K. A., Steingut, R. R., Vasquez, A. C., Yates, N., & Kennedy, A. A. (2019). Agency and high school science students' motivation, engagement, and classroom support experiences. *Journal of Applied Developmental Psychology*, 62, 77–92. <https://doi.org/10.1016/j.appdev.2019.01.004>

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences*, 13(3), 423–451. https://doi.org/10.1207/s15327809jls1303_6

Quenouille, M. H. (1949). On a method of trend elimination. *Biometrika*, 36(1–2), 75–91. <https://doi.org/10.2307/2332532>

Reeve, J. (2013). How students create motivationally supportive learning environments for themselves: The concept of agentic engagement. *Journal of Educational Psychology*, 105(3), 579–595. <https://psycnet.apa.org/doi/10.1037/a0032690>

Reeve, J., & Shin, S. H. (2020). How teachers can support students' agentic engagement. *Theory into Practice*, 59(2), 150–161. <https://doi.org/10.1080/00405841.2019.1702451>

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., & Lombardi, D. (2020). Evaluating sources of scientific evidence and claims in the post-truth era may require reappraising plausibility judgments. *Educational Psychologist*, 55(3), 120–131. <https://doi.org/10.1080/00461520.2020.1730181>

Smith, R. J. (2020). $P > 0.05$: The incorrect interpretation of "not significant" results is a significant problem. *American Journal of Physical Anthropology*, 172(4), 521–527. <https://doi.org/10.1002/ajpa.24092>

St. John, K. S., & McNeal, K. S. (2017). The strength of evidence pyramid: One approach for characterizing the strength of evidence of geoscience education research (GER) community claims. *Journal of Geoscience Education*, 65(4), 363–372. <https://doi.org/10.5408/17-264.1>

Tukey, J. W. (1958). A problem of Berkson, and minimum variance orderly estimators. *The Annals of Mathematical Statistics*, 29(2), 588–592. <https://doi.org/10.1214/aoms/117706637>

US Census. (2021). *Blinded for anonymity*.

van der Hoeven Kraft, K. J. (2017). Developing student interest: An overview of the research and implications for geoscience education research and teaching practice. *Journal of Geoscience Education*, 65(4), 594–603. <https://doi.org/10.5408/16-215.1>

Walsh, C., Quinn, K. N., Wieman, C., & Holmes, N. G. (2019). Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking. *Physical Review Physics Education Research*, 15(1), 010135. <https://doi.org/10.1103/PhysRevPhysEducRes.15.010135>

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician*, 73(sup1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>