Educational and Developmental Psychologist

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rdvp20

Climate crisis learning through scaffolded instructional tools

Janelle M. Bailey, Sonia Jamani, Timothy G. Klavon, Joshua Jaffe & Svetha Mohan

EDUCATIONAL AND DEVELOPMENTAL PSYCHOLOGIST
https://doi.org/10.1080/20590776.2021.1997065

Climate crisis learning through scaffolded instructional tools

Janelle M. Bailey [a], Sonia Jamani [b], Timothy G. Klavon [b*], Joshua Jaffe [b] and Svetha Mohan [b**]

[a]Department of Teaching and Learning, Temple University, Philadelphia, PA, USA; [b]Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, USA

ABSTRACT

Objective: Socially-relevant and controversial topics, such as the climate crisis, are subject to differences in the explanations that scientists and the public find plausible. Scaffolds can help students be evaluative of the validity of explanations based on evidence when addressing such topics and support knowledge gains.

Method: This study compared two scaffolds in which students weighed connections between lines of evidence and explanations for the topics of climate change and extreme weather events.

Results: A Wilcoxon-signed rank test showed that students' plausibility judgements shifted towards scientifically accepted explanations and that students increased their knowledge about climate crisis topics after completing both activities. A structural equation model suggested that students' shifts in plausibility judgements drive their knowledge gains for the extreme weather activity, but the climate change activity demonstrated a possible ceiling effect in its usefulness for learning.

Conclusions: When students choose their lines of evidence and explanatory models, their plausibility reappraisals result in greater levels of post-instructional knowledge. Although effect sizes were modest, the results of this study demonstrate that students' explicit reappraisal of plausibility judgements can support deeper learning of climate crisis issues.

ARTICLE HISTORY
Received 18 June 2021
Accepted 19 October 2021

KEYWORDS
Climate crisis; climate change; science education; scientific reasoning; argumentation; scaffold learning

KEY POINTS

What is already known about this topic:

(1) Students may have difficulty understanding complex or controversial topics such as the climate crisis.
(2) Evaluation is a key component of scientific thinking and a major piece of science and engineering practices.
(3) Instructional scaffolds can provide a way to help students learn how to evaluate competing models or explanations about a scientific phenomenon and lead to changes in their plausibility judgements about those explanations.

What this topic adds:

(1) Instructional scaffolds called Model-Evidence Link (MEL) diagrams have been created for two aspects of the climate crisis: the cause of climate change and the relationship between extreme weather and climate change.
(2) The two climate crisis MEL scaffolds support students in evaluating different explanations and learning about these topics.
(3) The build-a-MEL scaffold, where students choose aspects of the activity, resulted in greater changes in students' plausibility judgements than the preconstructed MEL.

Scientific reasoning is a key aspect of science learning and can be understood as an important skill involving the ability to judge the veracity of particular claims based on the quality of evidence provided. Socio-scientific topics, such as the climate crisis and its connections to extreme weather events (Zangori et al., 2017), are or will soon become pressing concerns for today's students (Sadler, 2004). As such, these topics may afford opportunities for students to hone their scientific reasoning abilities, particularly those involving the evaluation of competing claims (Sadler, 2004; Zangori et al., 2017). The climate crisis and extreme weather topics are pervasive themes in print and social media. Although the science behind these

CONTACT Janelle M. Bailey janelle.bailey@temple.edu
*Present address: School of Education, Black Hills State University.
**Present address: Department of Psychology, Tulane University.

phenomena – particularly climate change – is well established, the public, including K-12 students, are routinely exposed to a wide range of accurate, correct but misleading, and blatantly false claims, complicating their ability to effectively evaluate the credibility and plausibility of what they read or hear (Sinatra & Lombardi, 2020).

In order to teach socio-scientific topics, such as the climate crisis, some researchers are developing robust instructional tools to help scaffold students' scientific reasoning. Previously, Lombardi, Sinatra et al. (2013) compared such an activity with instructional materials developed for science education classes around the 1990s and 2000s (National Research Council [NRC], 1996). The present study compared two instructional scaffolds to examine students' evaluations of connections between scientific evidence and alternative explanations: the preconstructed Model-Evidence Link diagram activity (hereafter pcMEL; used by Lombardi, Sinatra et al., 2013) on climate change and the newer build-a-MEL (baMEL; Lombardi et al., 2020) on extreme weather events.

Background and theoretical framework

The climate crisis is a complex topic, and understanding it deeply involves the coordination of content-specific information around both large and small geographical areas and time scales (U.S. Global Change Research Program [USGCRP], 2009). Additionally, more general processes like evaluation and plausibility judgements are crucial to achieving deep understanding of this scientific phenomenon (Ford, 2015). Students' participation in scientific argumentation is dependent on their cognitive processes. In recent years, researchers have found that the reappraisal of plausibility judgements are important to help students experience conceptual change in science education (Lombardi, Nussbaum et al., 2016). Keeping this in mind, the current study focuses on scientific topics of climate change and extreme weather through scaffolded instructional methods that facilitate both evaluation and plausibility reappraisal.

In the following sections, we discuss the ideas of evaluation, plausibility judgements, and conceptual agency. These constructs are key components of the present study, with evaluation and plausibility judgements forming two of the variables of interest (the third being knowledge) and conceptual agency influencing the design of one of the two instructional scaffolds. After this, we describe scaffolding more generally followed by how it applies in the present study.

Evaluation

Evaluation is a core part of scientific learning and reasoning (Ford, 2015). It is crucial to evaluate scientific data and assess the evidence against proposed explanations. When students use the process of evaluation, they are replicating the authentic practices of scientists (National Research Council [NRC], 2012). As such, the process of evaluation is a key component of the scientific classroom learning environment (Ford, 2015; Lombardi, Sinatra et al., 2013). Additionally, researchers have found that students' evaluation of evidence and alternative models leads to stronger learning gains than when evaluating evidence against only a single explanation (Lombardi, Bailey et al., 2018).

Plausibility

Plausibility is a judgement that enables scientific inquiry and the creation and facilitation of scientific knowledge and reasoning (Lombardi, Nussbaum et al., 2016; Medrano et al., 2020). Plausibility judgements are tentative, informal evaluations of the potential truthfulness of an explanation (Lombardi, Nussbaum et al., 2016). A person can consider multiple explanations to be plausible simultaneously, or may change their judgements with new information. For example, a student could find the idea of humans influencing the climate to be highly plausible. At the same time, a competing explanation that changing output from the Sun could affect climate might be equally plausible to that student. Reading evidence that the Sun's output has decreased over the same time period that global temperatures have increased may cause the student to reappraise their original plausibility judgement. Lombardi, Nussbaum et al. (2016) argued that plausibility – and specifically plausibility reappraisal – influences the reconstruction of knowledge, especially in the area of science, supporting the idea of conceptual change. Providing students with opportunities to make both initial plausibility judgements and to reappraise those judgements in light of evidence can facilitate learning, particularly around complex or controversial socio-scientific topics.

Conceptual agency

Conceptual agency is when an individual identifies and uses resources to support their learning (Nussbaum & Asterhan, 2016; Pickering, 1995). Autonomy support, such as providing choice in learning, through curricular materials or instructional design has been shown to improve engagement in the learning process (Patall et al., 2018) and may be a more direct approach to supporting conceptual agency. Argumentation, which can be an autonomy-supportive strategy, can promote students' ability to understand and evaluate evidence via conceptual agency (Nussbaum & Asterhan, 2016), thus also providing opportunities for plausibility reappraisal. The conceptual dialogue that occurs when students are engaged in argumentation can affect students' learning and become transferable to other contexts (Nussbaum & Asterhan, 2016).

Scaffolded tools in an instructional setting

One commonly used metaphor within education is scaffolding. Often associated with Vygotsky's (1978) zone of proximal development, scaffolding refers to the support provided to students to help them achieve a task or learn a concept that may otherwise be out of their individual reach at the time (Reiser, 2004). Such support may come in the form of, for example, teacher assistance, curricular or instructional design, or peer interactions. Scaffolding is an effective teaching strategy to promote learning opportunities for students and assists in fostering students' scientific thinking and reasoning. Instructional scaffolds can prompt students to activate their prior knowledge and build new knowledge (Reiser, 2004), and can support the evaluation of claims and plausibility judgements.

The MEL instructional scaffold

Instructional treatments using graphical scaffolds have been tested through multi-year projects to examine the relationship between plausibility judgements and evaluation. The Model-Evidence Link (MEL) diagram activities are scaffolds that facilitate students in making connections between evidence and models (Chinn & Buckland, 2012; Lombardi, 2016; Lombardi, Sibley et al., 2013). MEL diagrams support student understanding of core science education content via scientific and engineering practices (NRC, 2012). Each MEL activity includes scientific evidence along with alternative models (i.e., explanations) that promote students' evaluations, creating student engagement in the science classrooms. Additionally, the build-a-MEL scaffold in particular provides autonomy support by allowing students to choose different models and lines of evidence that they will subsequently evaluate.

In previous studies, high school students' plausibility reappraisals and evaluation related to shifts in their knowledge for topics like climate change (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018). Previous studies also showed that the MEL motivated students and promoted scientific thinking and learning (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Medrano et al., 2020). In this project, the MEL diagram allows students to evaluate connections between lines of scientific evidence and two or three alternative and competing explanations of a phenomenon (e.g., a scientific and one or two non-scientific alternative explanations about the causes of current climate change or the relationship between extreme weather and climate change). The MEL activities intentionally require that students create more tangible reasoning by helping students evaluate the relationship between scientific evidence and these competing explanations. Although previous research has shown promise for the MEL activities, it is possible that the scaffolds may be more or less effective for differing topics. Thus, continued research focusing on the different topics covered by MEL and baMEL activities is warranted. Additionally, with further studies we hope to better understand the role of autonomy support, and by extension conceptual agency, provided by the baMELs in particular.

The current study

The current study compared two scaffolds to examine students' evaluations of connections between scientific evidence and alternative explanations: the pcMEL and baMEL activities. In the Climate Change (CC) pcMEL, students were presented four lines of scientific evidence and two models about causes of current climate change. In the Extreme Weather (EW) baMEL, students investigated the relationship between extreme weather events and climate change through constructing their own MEL diagrams by selecting four lines of evidence from eight choices and two alternative explanatory models from three. By building their own models, we hypothesized that the baMEL activity could promote deeper conceptual agency (Nussbaum & Asterhan, 2016) and result in greater knowledge gains. The autonomy supportive baMEL would facilitate students' authority to collectively reason in their classroom-based learning environments by facilitating students' evaluations of the connections between scientific evidence and alternative models about issues relating to the climate crisis. This would be related to stronger shifts in plausibility judgements towards the scientific explanation and greater knowledge gains than for the pcMEL. Thus, we examined the following research questions:

(1) How do students' plausibility judgements and knowledge change over the course of each of two instructional treatments (pcMEL and baMEL)?
(2) To what extent do students' levels of evaluation influence students' plausibility shifts and knowledge gains in each MEL topic (climate change + extreme weather)?
(3) How does the relation between students' evaluations, plausibility judgements, and knowledge compare between the pcMEL and baMEL?

Hypotheses

Research question 1

The MEL and baMEL scaffolds, broadly, have each shown promise for improving students' plausibility judgements and knowledge (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Medrano et al., 2020). We believe the same will be true for the MEL and baMEL scaffolds relating to the climate crisis used in this study. In the case of plausibility judgements, we consider it to be an improvement when students rate the scientific model as being more plausible than one or more alternative models, or to increase the plausibility gap (i.e., difference) between them if the scientific model is already rated higher. Knowledge shifts are considered improving when the students' ratings of provided statements are more aligned with what climate scientists would say. Both the plausibility judgement and knowledge variables and their measurement are described in greater detail in the Method section below.

Research questions 2 & 3

Our second and third questions focus on the relationships between students' levels of evaluation, plausibility shifts, and knowledge gains, analysed using partial least squares structural equation modelling (PLS-SEM). We will construct a preliminary model for each topic, the CC pcMEL and the EW baMEL (Figure 1). These hypothetical models are based on the theoretical perspective around plausibility by Lombardi, Nussbaum et al. (2016) as well as the relationships developed in our previous research (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018). The model is centred around the students' level of evaluation with pre instruction knowledge and plausibility gap as antecedent indicators and post instruction plausibility gap, followed by post instruction knowledge, as subsequent indicators. We expect students' levels of evaluation to have a strong relationship with their post instruction plausibility gap and, in turn, with their knowledge gains. We also expect the EW baMEL to show a stronger relationship between the post instruction plausibility gap and the students' post instruction knowledge gains than in the CC pcMEL.

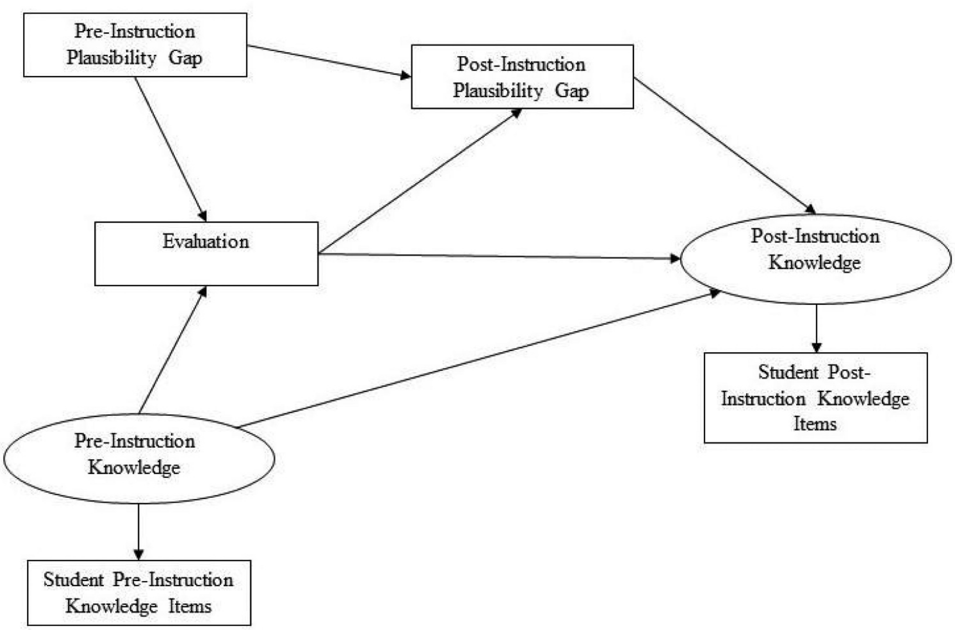

Figure 1. Initial PLS-structural equation model relating plausibility, evaluation, and knowledge. Indicators (i.e., observed values) are designated by rectangles and constructs (i.e., derived values) are designated by ovals. The Climate Change pcMEL knowledge score consisted of 4 items. The Extreme Weather baMEL knowledge score consisted of 10 items.

Method

Participants

Over three hundred (*N* = 313) middle school (Grades 6-8) and high school (Grades 9-12) students enrolled in Earth and environmental science classrooms in the Mid-Atlantic and Southeastern regions of the U.S. participated in the study. Of the 313 students who provided assent and parental consent, 88 middle school and 83 high school students completed all activities and were thus included in the final analysis (*n* = 171); reasons for this reduction in sample size are discussed further in the Results section. There were no significant differences between the middle school and high school students' knowledge scores at pre instruction, and thus the two groups were analysed together. Although self-report demographic data were collected from many of the participants, not everyone completed this request. The participating students' demographics were generally representative of their schools (Table 1). The school demographics show some variation across race/ethnicity and socioeconomic status, but past research has shown that there have not been differences between schools showing such variation (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018).

Instructional topics and scaffolds

In the current study, the intervention entailed two scientific topics, climate change and extreme weather, addressed in two MEL scaffolds. Both of these topics were part of the curriculum in the participating middle school and high school classrooms. These topics are also connected to Next Generation Science Standards (NGSS Lead States, 2013) and state standards for these school districts. The topic of climate change was the basis for the pcMEL, while the topic of extreme weather was the content for the baMEL (Figure 2).

Climate change pcMEL

The CC pcMEL addresses the cause of current climate change (e.g., Lockwood, 2009; Oreskes, 2004; Table 2). Discussions of evidence include, for example, tracking emissions of greenhouse gases or solar output over long periods of time. A one-page evidence text document, aimed at approximately the eighth-grade reading level, elaborates on each line provided in Table 2. Each evidence text also presents one or more figures, graphs, or tables to help students further engage with these common tools of scientific writing. Sources for information in the CC pcMEL models and lines of evidence include scientific journals (e.g., *Science* and *Journal of Geophysical Research: Atmospheres*) and government sites such as NOAA and NASA.

In the first part of the CC pcMEL, students were introduced to two models (the scientifically accepted explanation and a non-scientific alternative) and four lines of scientific evidence. Students completed the activity at or near the beginning of an instructional unit on weather and climate. Students evaluated the scientific evidence and models to make a rational judgement about the connections between them. Participants were instructed to draw different types of arrows from each line of evidence to each model based on how they thought the evidence relates to the model. Four different types of arrows were used: a squiggly arrow indicated the participant believes that the evidence strongly supported 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 (Figure 2(a)).

Extreme weather

The EW baMEL focuses on weather-related events. Some examples of the evidence include the number of times of yearly rainfall in the U.S. during the 20th century, the amount of increase in North Atlantic

Table 1. Demographic characteristics of participating schools.

School	Teacher	*n*	Sex	Race (%)	Economic Composition
MA1[a]	06	88	Female 49.1%	White 70.1%, Hispanic 17.6%, Black 5%, Asian 3.7%, Native Hawaiian/Pacific Islander 0.1%, Two or more races 3.3%	Economically disadvantaged students 29.1%
MA2[a]	0107	813	Female 47.6%	White 52.3%, Hispanic 35.1%, Black 8.2%, Asian 4.5%, Native Hawaiian/Pacific Islander 0.1%. Two or more races 0.3%	Economically disadvantaged students 27.0%
SE1[b]	525560	32534	Female 46.2%	White 79.1%, Hispanic 14.3%, Black 3.5%, Asian 1.8%, American Indian/Alaska Native 0.5%, Two or more races 3.4%	Economically disadvantaged students 19.2%

MA1 = data from Mid-Atlantic School 1, MA2 = data from Mid-Atlantic School 2, and SE 1 = data from Southeast School 1.
[a][MA] Department of Education (2018-2019).
[b][SE] Department of Education (2018-2019).

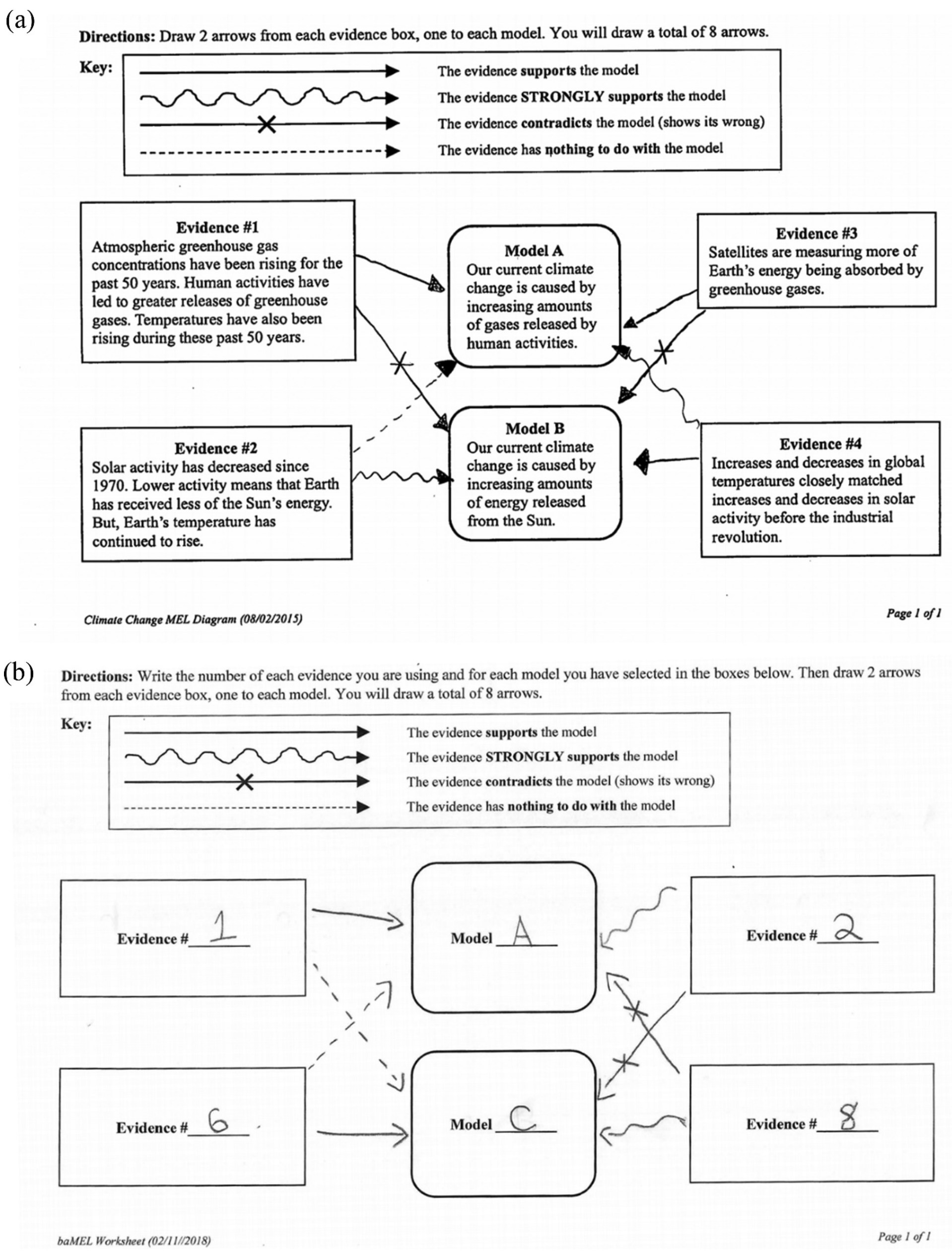

Figure 2. Student examples of the (a) climate change pcMEL diagram and (b) extreme weather baMEL diagram.

Table 2. Models and lines of evidence in the climate change preconstructed MEL (CC pcMEL).

Label	Statement
Model A	Climate change is caused by humans who are releasing gases into the atmosphere.
Model B	Climate change is caused by increasing amounts of energy released from the Sun.
Evidence #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.
Evidence #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.
Evidence #3	Satellites are measuring more of Earth's energy being absorbed by greenhouse gases.
Evidence #4	Increases and decreases in global temperatures closely matched increases and decreases in solar activity before the industrial revolution.

tropical storm power since 1970, and the amount of European snowfall over the past decade. These are presented as eight lines of scientific evidence (Table 3). The development team synthesized the findings and the lines of evidence based scientific journals (e.g., *The Bulletin of the American Meteorological Society*

Table 3. Models and lines of evidence in the extreme weather build-a-MEL (EW baMEL).

Label	Statement
Model A	The number and strength of extreme weather events vary naturally. Human activities release carbon in the atmosphere. Yet, plants and oceans absorb any carbon increases.
Model B	Increases in extreme weather events are linked to climate change. Current climate change is mainly caused by human activities, such as fossil fuel use.
Model C	Over time, increases and decreases in extreme weather events are mainly caused by changes in Earth's orbit around the Sun.
Evidence #1	Since 1950, Earth's atmosphere and oceans have changed. The amount of carbon released to the atmosphere has risen. Dissolved carbon in the ocean has also risen. More carbon has increased ocean acidity and coral bleaching.
Evidence #2	From 1910 to 1995, record rainfall events increased across the United States. Over the same time period, there was a sharp increase in the amount of carbon released to the air. Much of this carbon comes from fossil fuel use.
Evidence #3	Ocean sea surface temperatures have increased since about 1970. In the North Atlantic, tropical storm power has also increased over this same time period. A storm's power depends on its strength and how long it lasts.
Evidence #4	Since 2000, there have been more intense, extreme, weather events around the world. Record rainfall fell in Europe. The southeastern United States had the most active month of tornadoes. The decade from 2000 to 2010 was the warmest ever during the past 1000 years.
Evidence #5	Frequency and size of large wildfires have increased in the Western U.S. since 1970. Average spring and summer temperatures have also risen in the Western U.S. during this time.
Evidence #6	In the last 100 years, global temperatures have increased. In that same time period, heavy precipitation events have also increased.
Evidence #7	Arctic Ocean sea ice extent has declined, with the Arctic warming at a pace two to three times the planet's average. Over the last decade, record cold temperatures and snowfall have occurred in Europe and Asia.
Evidence #8	Earth's orbit is elliptical. But, the shape of the ellipse is almost a perfect circle. In the Northern Hemisphere, Earth is slightly closer to the Sun in winter than in summer.

and *Nature*) and government sites such as NOAA. Furthermore, a one-page evidence text approximately at the eighth grade reading level is provided for each line of evidence. Three models (one scientific and two alternative) are also presented.

Participants chose four out of the eight lines of scientific evidence and two of the three models to construct their own diagram. Unlike the CC pcMEL diagram, where models were provided, the EW baMEL diagram contained blank boxes (Figure 2(b)). In these boxes, participants wrote which lines of evidence and models they selected. Similar to the CC pcMEL, participants connected the evidence texts to models using the four different types of arrows, drawing eight total arrows.

Evaluation/explanation task

We compared evaluation for both scaffolds, the CC pcMEL and EW baMEL, using participants' written responses detailing their evidence-to-model link connections. Participants selected one or two of the connections that they made and explained their evaluation of that particular line of evidence and model connection. A rubric, originally developed by Lombardi, Brandt et al. (2016), was used to score each explanation. Coders rated explanations on different levels of evaluation, where: 1 = Erroneous, 2 = Descriptive, 3 = Relational, and 4 = Critical. These categories established well-defined levels of evaluation to reflect accuracy and explanations present in the participants' responses. To establish scoring reliability, two authors independently scored 25% of the explanation task responses using the evaluation score rubric. The two raters' scores had an intraclass correlation coefficient (ICC) of 0.69, indicating an acceptable level of reliability between coders, with full consensus met after discussion. All raters received the same training to evaluate responses using the rubric. The initial rating may have been a function of the raters having different levels of experience with the rubric, with greater agreement occurring over time. One coder completed scoring for the remainder of the responses.

Plausibility judgements

For both the CC pcMEL and EW baMEL, students were instructed to rate the plausibility of all explanatory models at pre and post instruction. In the case of the EW baMEL, students recorded their plausibility judgements for all three explanatory models at both points, even though they only worked with two of the models on the diagram. Students gauged the plausibility of each model using a 1–10 scale (where 1 = greatly implausible and 10 = highly plausible), based on previous measures used with MEL activities (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Lombardi, Sinatra et al., 2013). Because the CC pcMEL offered two explanatory models, plausibility gap scores were calculated as the rating of the scientific model minus the alternative model. The EW baMEL offered three different explanatory models (scientifically accepted and two alternative options), but in order to match the scoring from the CC pcMEL, plausibility gap scores were calculated as the rating of the scientific model minus the average of the two alternative models. This is reasonable given that there were no significant differences between the rating of the two alternatives at either pre ($t = 1.609$, $p = .109$) or post instruction ($t = -0.026$, $p = .980$). Plausibility gap scores ranged from −9 to +9, where positive scores indicated that participants judged the scientific model as more plausible than the alternative model(s), and negative scores indicated that participants judged the alternative model(s) as being more plausible than the scientific model.

Knowledge

For both the CC pcMEL and EW baMEL, participants completed a multi-item knowledge survey at pre and post instruction. The CC knowledge survey contained five items and the EW knowledge survey contained eleven items. Students rated the degree to which they think *climate scientists* would agree or disagree with each statement, with ratings on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). Questions were constructed in both positive and negative orientations (i.e., in effect scientists would disagree with these knowledge statements) and we reverse coded these negatively worded statements for analysis. Knowledge scores were the average of the students' ratings after reverse coding.

Procedures

During the summer, middle school and high school teachers participated in a three-day professional development workshop with the project team. The workshops focused on introducing and practicing using the pcMEL and baMEL instructional scaffold activities. The goal was to cover the content and pedagogical strategies for effective classroom implementation using the MEL activities. As a whole, participating teachers agreed to introduce each MEL activity at or near the beginning of the unit relating to weather and climate. The CC pcMEL activities were taught first and the EW baMEL activities were taught second. Each activity took approximately 90 minutes, spread out over multiple class meetings. The lessons may not have been taught consecutively but rather may have included other lessons between the two MELs.

Prior to the CC pcMEL activity, students performed a plausibility ranking task, which served as an introduction to the ideas of plausibility, falsifiability, and critical evaluation. Students ranked the importance of different types of evidence for determining the plausibility of an explanatory model: 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 – 4, they read a small passage on falsifiability to evaluate the role of scientific evidence and re-ranked the items. Next, for a given treatment (i.e., pcMEL or baMEL), students completed the knowledge survey and model plausibility ratings (pre) for each explanatory model on that topic.

When completing the CC 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 the CC knowledge survey (post). When completing the EW baMEL activity, they were introduced to the three alternative models explaining the topic. Students next read the texts for all eight lines of evidence. Thereafter, small groups of students worked together to select four lines of evidence from the eight available and two alternative models from the three available. These students used these selected lines of evidence and models to construct a MEL diagram, which they then completed (i.e., by drawing arrows) in the same manner as the CC pcMEL. Similar to the CC pcMEL activity, students then worked individually to complete the explanation task. The activity ended with the second iteration of the model plausibility ratings and EW knowledge survey.

Results

In this section, we present the results of our study in relation to our research questions. We start with general data screening analyses then move to each research question in turn. Data screening included the removal of any participating students whose data were incomplete (leading to a reduction from $N = 313$ to $n = 171$). All students in the classes participated in the multi-day lesson, but some may be missing data due to absence, disengagement in the activities, or lack of completion of one or more pieces of the activities for other individual reasons. There were no systematic issues of missing data.

We imported data into JASP 0.14.1 (JASP Team, 2020) for analysis, then looked at descriptive statistics for each of the variables (evaluation, plausibility, and knowledge). Most values of skewness and kurtosis were acceptable (i.e., less than |1|; George & Mallery, 2009). Those that were outside of the acceptable range, but still less than |2|, include the average explanation score for the EW baMEL, for which participants generally had quite low scores, and the CC pcMEL knowledge scores (both pre and post), for which students generally had higher scores. This was one reason that we used nonparametric statistical analyses as described below (Nussbaum, 2015).

Reliability

We ran a reliability analysis using McDonald's ω for the two knowledge instruments (evaluation and plausibility are single-measurement variables). The CC knowledge instrument had one item with a negative correlation, likely due to a ceiling effect on the question. This item was removed and the remaining items had a reliability of ω = 0.419 at pre-test and ω = 0.549 at post-test. Similarly, the EW knowledge instrument showed negative correlations between three items and the scale at the pre-test and between one item and the scale at post-test. After inspection, one of the items (the one negatively correlated on both pre and post-test) was removed but the other two were retained due to their strong relation to the content of the activity. The reliability of the resulting instrument was ω = 0.301 at pre-test and ω = 0.466 at post-test. These values are all lower than are generally considered acceptable, but typical measures of reliability are often sensitive to the homogeneity of the sample (Thompson, 2003). Thus, it may be that the students in the sample have a similar level of understanding on these measures.

RQ 1: repeated measures comparisons

Our first research question asks:

(1) How do students' plausibility judgements and knowledge change over the course of each of two instructional treatments (pcMEL and baMEL)?

We used a nonparametric repeated measures comparison (Wilcoxon signed-rank test) to answer this question for both plausibility and knowledge as described in the next two sections.

Plausibility

We ran repeated measures comparisons of the plausibility scores, looking for changes over time (i.e., pre to post) for each instrument (Figure 3). The CC pcMEL plausibility score at pre-test was M = 1.16 (SD = 3.76) and at post-test was M = 2.00 (SD = 3.76). The EW baMEL plausibility score at pre-test was M = 1.71 (SD = 3.33) and at post-test M = 3.08 (SD = 2.98). The Wilcoxon statistics showed significant increases from pre to post for both the CC pcMEL (W = 2877.5, $p < .001$, rank-biserial correlation = 0.354, moderate effect; McGrath & Meyer, 2006) and the EW baMEL (W = 3703.0, $p < .001$, rank-biserial correlation = 0.395, large effect).

Knowledge

We also looked at the changes from pre to post of the knowledge scores for each scaffold (Figure 4). Means and standard deviations of the four knowledge measures were: CC pcMEL pre M = 3.89 (SD = 0.56); CC pcMEL post M = 4.14 (SD = 0.58); EW baMEL pre M = 3.39 (SD = 0.37); and EW baMEL post M = 3.58 (SD = 0.44). The repeated measures Wilcoxon statistic again showed statistically significant increases in scores for both the CC and EW knowledge instruments (CC: W = 1978.5, $p < .001$, rank-biserial correlation = 0.563, large effect; EW: W = 2573.5, $p < .001$, rank-biserial correlation = 0.533, large effect; McGrath & Meyer, 2006).

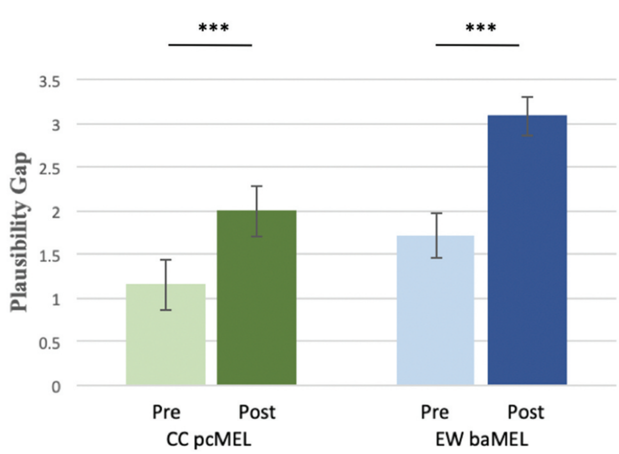

Figure 3. Plausibility gap scores for each instructional treatment. Range 1 (highly implausible)+9 (highly plausible). Errors bars indicate ±1 standard error. Asterisks indicate a statistically significant difference between pre- and post-instruction, ***$p < 0.01$.

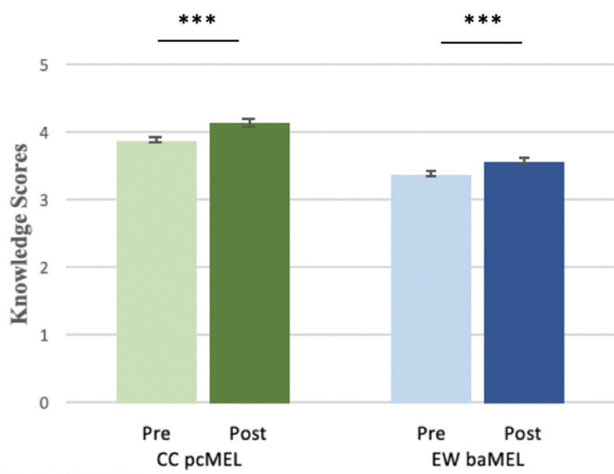

Figure 4. Knowledge scores for each instructional treatment. Range 1 (strongly disagree with scientific)–5 (strongly agree with scientific). Error bars indicate ±1 standard error. Asterisks indicate a statistically significant difference between pre- and post-instruction, ***p < 0.001.

RQs 2 and 3: structural equation modelling

We used a multi-faceted approach when analysing the relationships between the variables present in the MEL-diagram activities to answer our second and third research questions:

(2) To what extent do students' levels of evaluation influence students' plausibility shift and knowledge gains in each MEL topic (climate change + extreme weather)?
(3) How does the relation between students' evaluations, plausibility judgements, and knowledge compare between the pcMEL and baMEL?

After organizing the data, we analysed it using WarpPLS 7.0 (Kock, 2020). Our choice of partial least squares structural equation modelling (PLS-SEM) is ideal for the small sample set as it uses ranked data and is distribution free (Lombardi, Danielson et al., 2016). The use of PLS-SEM in this fashion has come under attack (Goodhue et al., 2012), though Kock (2020) indicated that Goodhue et al.'s (2012) use of low path coefficients for small and medium effect sizes may have exacerbated any negative effects found in their test simulations. Additionally, we employed jackknifing as the resampling technique for PLS-SEM. Jackknifing reduces standard error and may increase statistical power by removing one or more indicators at a time and replacing them with partial estimates (Abdi & Williams, 2010; Quenouille, 1949; Tukey, 1958). This replacement seeks to increase the predictive ability of the PLS-SEM (Kock, 2020). We made model comparisons using Tenenhaus Goodness of Fit (GoF), which answers how well different subsets of the data can be explained by the model (Henseler & Sarstedt, 2013).

After completing the PLS-SEM, we implemented a holistic approach to evaluating the relationships formed by the model, using the significance, the beta weight, and the effect size of each link. Though significance (i.e., p-value) plays an important role in how we assess our data, there are arguments that -p – value alone should not exclude relationships in the light of strength of the connection (i.e., beta weight) or the effect size (i.e., importance as measured by Cohen's f-squared; Smith, 2019). Wasserstein et al. (2019) implore us to not "believe that an association or effect is absent just because it was not statistically significant" (p. 1). This holistic approach provides us the opportunity to understand the relationships between the variables in ways that may help us provide students with tools to increase their levels of evaluation and knowledge gains with these instructional scaffolds.

Climate change pcMEL PLS-SEM

Upon running the PLS-SEM software, we found that the CC pcMEL produced a large goodness of fit (Tenenhaus GoF = 0.475, large ≥ 0.36; Wetzels et al., 2009). The use of Tenenhaus GoF is warranted when implementing PLS-SEM, as PLS-SEM does not optimize any global scalar function (Tenenhaus et al., 2005). Therefore, it lacks a global validation index, such as

Chi-squared for maximum likelihood modelling, to provide user insight into the validity of the model. In the case of PLS-SEM, GoF serves the same function *Chi-squared* or the root mean square error of approximation.

This model produced limited (Table 4) results as it appears that students' pre instruction knowledge (PrK) drives most of the post instruction knowledge (PoK) gains (β = 0.62, $p < .001$, $f^2 = 0.402$). PrK also shows a small effect on students' levels of critical evaluation (E; β = 0.22, $p = .03$, $f^2 = 0.048$). Pre instruction plausibility gap (PrP) exhibited a similar relationship with post instruction plausibility gap (PoP; β = 0.61, $p < .001$, $f^2 = 0.385$). As the pathways between knowledge and evaluation were connected to those of plausibility by only weak relationships, the creation of the model was deemed unsuccessful and we did not produce a final image for it. We will propose causes and implications for these relationships in the Discussion section.

Table 4. PLS-structural equation modelling β weights, effect sizes, and significance values for the climate change pcMEL relationships.

	Pre-Instruction Plausibility			Evaluation			Post-Instruction Plausibility			Pre-Instruction Knowledge		
	β	f^2	p	β	f^2	p	β	f^2	p	β	f^2	p
Evaluation	−0.06	0.003	.47	-	-	-	-	-	-	0.23	0.048	.03
Post-Instruction Plausibility	0.61	0.385	.000	0.09	0.017	.20	-	-	-	-	-	-
Post-Instruction Knowledge	-	-	-	0.14	0.033	.06	0.05	0.012	.37	0.62	0.402	.000

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

Table 5. PLS-structural equation modelling β weights, effect sizes, and significance values for the extreme weather baMEL relationships.

	Pre-Instruction Plausibility			Evaluation			Post-Instruction Plausibility			Pre-Instruction Knowledge		
	β	f^2	p	β	f^2	p	β	f^2	p	β	f^2	p
Evaluation	−0.18	0.037	.43	-	-	-	-	-	-	0.23	0.056	.02
Post-Instruction Plausibility	0.49	0.260	.39	0.15	0.091	.055	-	-	-	-	-	-
Post-Instruction Knowledge	-	-	-	0.11	0.513	.41	0.42	0.229	.000	0.26	0.221	.12

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

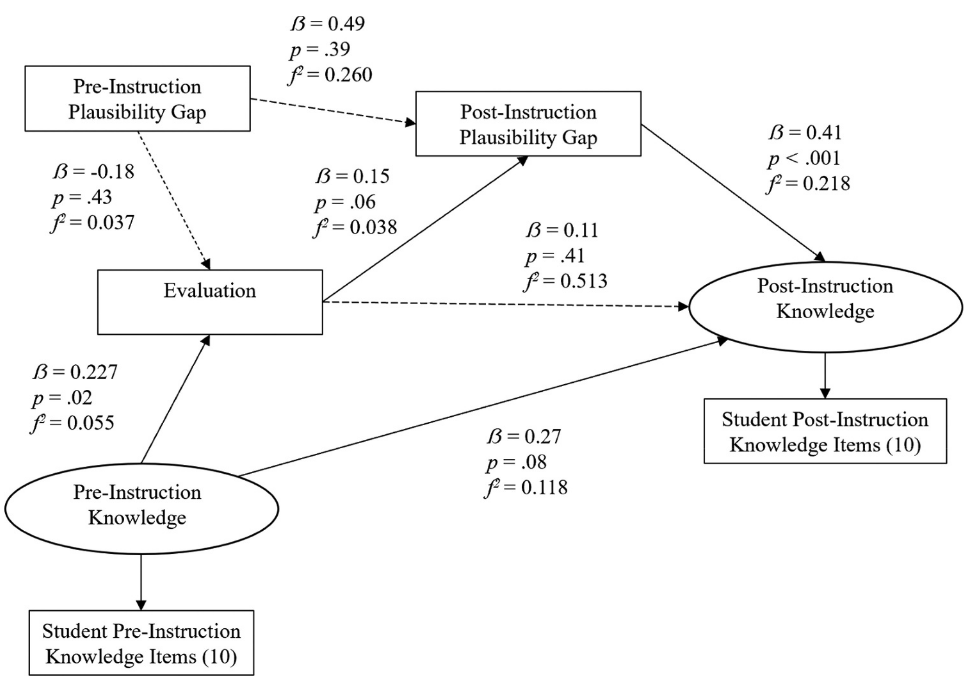

Figure 5. PLS-structural equation model relating plausibility, evaluation, and knowledge in the extreme weather baMEL. Indicators (i.e., observed values) are designated by rectangles and constructs (i.e., derived values) are designated by ovals. The Extreme Weather knowledge score consisted of 10 items.

Extreme weather baMEL PLS-SEM

The EW baMEL data produced a PLS-SEM with a large goodness of fit (Tenenhaus GoF = 0.410; Table 5 and Figure 5), indicating that this model is also highly representative of the data. In this model, the influence of PrK on both PoK (β = 0.27, p = .08, f^2 = 0.118) and E (β = 0.227, p = .02, f^2 = 0.055) was small. The E-PoP relationship was also small (β = 0.15, p = .06, f^2 = 0.038). However, the PoP-PoK relationship appears to be more important, with a stronger relationship and medium effect size (β = 0.41, p < .001, f^2 = 0.218).

Discussion

The current study revealed that both MEL scaffolds are effective for learning, however the results only partially supported our hypothesis that the baMEL would be more effective than the pcMEL. In the following sections, we discuss each research question in more detail.

RQ 1: repeated measures comparisons

Students underwent shifts in their plausibility judgements that moved towards the scientific explanation from pre to post instruction on both the CC pcMEL and the EW baMEL. In looking at the plausibility ratings (i.e., raw scores rather than plausibility gaps or differences), we see that for the CC pcMEL, the mean plausibility rating of the scientific model rose slightly (from 7.28 pre to 7.79 post, on a scale of 1-10) but that the mean plausibility rating of the alternative model decreased (from 6.12 pre to 5.79 post). Likewise, the plausibility rating of the EW baMEL's scientific model increased (7.03 pre to 8.14 post) while those of the alternative models decreased (model A: 5.50 pre to 5.05 post; model C: 5.14 pre to 5.06 post). Thus, changes to the plausibility gap were a function of both increased plausibility of the scientific model and decreased plausibility of alternative explanations. The plausibility reappraisals were particularly helpful in the case of the EW baMEL, as discussed further below, showcasing students' ability to engage with scientific inquiry and reasoning. We also looked at the changes from pre to post of the knowledge scores for each scaffold and found that both had statistically significant improvements with large effect sizes, indicating that students improved their knowledge over the course of each MEL activity.

RQs 2 and 3: structural equation modelling

RQ 2 asked how the plausibility, knowledge, and evaluation scores related to each other within each of the MEL activities. The PLS-SEM analysis performed on the CC pcMEL was not fruitful. The CC pcMEL has been used for nearly 10 years (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Lombardi, Sinatra et al., 2013) and, over that time, has potentially begun to show a "ceiling effect" in its effectiveness with secondary science students. Aksit et al. (2017) speculate that current students are growing up in a time where climate change has greater scientific certainty, increased consensus about the topic, and more exposure during formal and informal educational experiences. These conditions might also be impacting the effectiveness of scaffolds about climate change such as the MEL activity.

The internal relationships for the EW baMEL were much more robust than those for the CC pcMEL. The post instruction plausibility reappraisal was the strongest influence of student knowledge gains within this activity. Additionally, the pre to post instructional knowledge link is weaker than much of our previous research (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Medrano et al., 2020). This leads us to surmise that it is the students' reappraisal of their plausibility that is driving their gains in this activity (Lombardi, Nussbaum et al., 2016).

To answer RQ 3, we turn to the very different outcomes of the PLS-SEM analyses for each model, which are indicative of the relationships between the factors across the instruments. The pre to post plausibility and pre to post knowledge relationships in the CC pcMEL were independent of each other (i.e., linked by weak relationships), whereas in the EW baMEL, the students' reappraisal of their plausibility judgements appears to be driving their knowledge gains.

Wasserstein et al. (2019) remind us that all statistical analysis is influenced by the "expert judgment" (p. 5) of the researcher performing it. It is notable for this project that after analysis, each of the strong relationships was also statistically significant. As Wasserstein and colleagues also encourage researchers towards openness, we believe it is important for us to be thorough in our communication about our approach to decision-making. During this analysis, we recognize that previous research (Lombardi, Bailey et al., 2018; Lombardi, Bickel et al., 2018; Lombardi, Sinatra et al., 2013; Medrano et al., 2020) also supported that these instruments provide reliable findings of plausibility shifts towards scientific models and knowledge gains among students.

Limitations

Studies conducted in the classroom pose limitations and challenges. For instance, as noted by the reduction of our sample size by close to half, student attendance and completion rates in a multi-day study can complicate data collection. Additionally, each student has their own experience and prior knowledge with each relevant topic (Lawson, 1988). While we attempt to account for this through our modelling, any teacher can attest to the in-the-moment challenges of pacing when students have different starting points. Another consideration is student engagement. In a science classroom, it is important to keep the student engaged in order to be able to link prior information to the new information (Sinatra et al., 2015). We believe, through both completion of the written documents and observations in classrooms using the MEL activities, that engagement is generally quite good but may still not be as high as one would like. A final limitation may be the aforementioned ceiling effect for the topic of climate change in the classroom, rendering the CC pcMEL less effective than in previous studies (e.g., Lombardi, Sinatra et al., 2013).

Implications for research and teaching

While the MEL has been informative in the past, the CC pcMEL specifically may not be as useful as a research instrument in the current time due to a ceiling effect. We believe, however, that the CC pcMEL still has value in the classroom setting, as it can provide an introduction to the MEL activities in general and specifically to the more robust EW baMEL. It is also crucial to see how science students' evaluation skills improve (or do not) over the long term and with multiple MEL use through ongoing studies. In terms of teaching, the current study has been a part of a multi-year project that demonstrates multi-tier scaffolds with related topics (e.g., CC pcMEL and EW baMEL) can be useful for learning socio-scientific issues. Scaffolding techniques may be beneficial for teachers to use in the science classrooms to create more student engagement and curiosity towards science related topics.

Conclusion

The results of the study revealed that the CC pcMEL and EW baMEL are both effective for learning. Both MEL scaffolds promoted plausibility shifts towards the scientific model and increased knowledge in the socio-scientific topic of climate change and its effect on extreme weather events. The MEL scaffolds show to be an effective tool for promoting scientific literacy, playing a positive role in helping students improve their knowledge and can even help them solve some of the local and global socio-scientific problems like the climate crisis. As we continue our research project, we will refine and test these more robust scaffolds that facilitate students' conceptual agency and help them prepare to think scientifically.

Acknowledgments

Sonia Jamani and Janelle Bailey worked in tandem as co-lead authors and shared responsibilities equally, with authorship listed alphabetically. The team would like to thank the special issue editor, Doug Lombardi, for his valuable insights into the project. We also appreciate the support of the Science Learning Research Group.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the US National Science Foundation (NSF) under Grant Nos. 1721041 and 2027376. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the NSF's views.

ORCID

Janelle M. Bailey http://orcid.org/0000-0001-9563-2016
Sonia Jamani http://orcid.org/0000-0002-4833-0370
Timothy G. Klavon http://orcid.org/0000-0002-2890-0970
Joshua Jaffe http://orcid.org/0000-0002-2715-0215
Svetha Mohan http://orcid.org/0000-0002-1234-4244

Data availability statement

The data that support the findings of this study are available from the corresponding author, Janelle Bailey, upon reasonable request.

References

Abdi, H., & Williams, L. J. (2010). Jackknife. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 655–660). Sage.

Aksit, O., McNeal, K. S., Gold, A. U., Libarkin, J. C., & Harris, S. (2017). The influence of instruction, prior knowledge, and values on climate change risk perception among undergraduates. *Journal of Research in Science Teaching*, *55*(4), 550–572. https://doi.org/10.1002/tea.21430

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. K. Brem, & G. M. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning about evolution* (pp. 211–232). Oxford University Press.

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

George, D., & Mallery, P. (2009). *SPSS for windows step by step: A simple guide and reference, 16.0 update* (9th ed.). Pearson Education.

Goodhue, D. L., Lewis, W., & Thompson, R. (2012). Does PLS have advantages for small sample size or non-normal data? *MIS Quarterly*, *36*(3), 981–1001. https://doi.org/10.2307/41703490

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

JASP Team. (2020). *JASP* (Version 0.14.1) [Computer software].

Kock, N. (2020). *WarpPLS user manual* (Version 7.0) [Computer software]. ScriptWarp Systems.

Lawson, A. E. (1988). The acquisition of biological knowledge during childhood: Cognitive conflict or tabula rasa? *Journal of Research in Science Teaching*, *25*(3), 185–199. https://doi.org/10.1002/tea.3660250304

Lockwood, M. (2009). Solar change and climate: An update in the light of the current exceptional solar minimum. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *466*(2114), 303–329. https://doi.org/10.1098/rspa.2009.0519

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., 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., Sibley, B., & Carroll, K. (2013). What's the alternative? Using model-evidence link diagrams to weigh alternative models in argumentation. *The Science Teacher*, *80*(5), 50-–55.

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

Lombardi, D., Uslu, B., & Bailey, J. M. (2020). Extreme weather events and the climate crisis: What is the connection? *The Earth Scientist*, *36*(3), 22–26. https://www.nestanet.org/cms/sites/default/files/journal/Fall20.pdf

Lombardi, D. (2016). Beyond the controversy: Instructional scaffolds to promote critical evaluation and understanding of Earth science. *The Earth Scientist*, *32*(2), 5–10. https://www.nestanet.org/resources/Documents/Advocacy/TES/2015-2020/Summer16.pdf .

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods*, *11*(4), 386–401. https://doi.org/10.1037/1082-989X.11.4.386

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 Research Council. (1996). *National science education standards*. National Academy Press. https://doi.org/10.17226/4962

National Research Council. (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/

Nussbaum, E. M., & Asterhan, C. S. C. (2016). The psychology of far transfer from classroom argumentation. In F. Paglieri (Ed.), *The psychology of argument: Cognitive approaches to argumentation and persuasion* (pp. 407–423). College Publications.

Nussbaum, E. M. (2015). *Categorical and nonparametric data analysis: Choosing the best statistical technique*. Routledge.

Oreskes, N. (2004). The scientific consensus on climate change. *Science*, *306*(5702), 1686–1686. https://doi.org/10.1126/science.1103618

Patall, E. A., Hooper, S., Vasquez, A. C., Pituch, K. A., & Steingut, R. R. (2018). Science class is too hard: Perceived difficulty, disengagement, and the role of teacher autonomy support from a daily diary perspective. *Learning and Instruction*, *58*, 220–231. https://doi.org/10.1016/j.learninstruc.2018.07.004

Pickering, A. (1995). *The mangle of practice: Time, agency, and science*. The University of Chicago Press.

Quenouille, M. H. (1949). On a method of trend elimination. *Biometrika*, *36*(1/2), 75–91. https://doi.org/10.1093/biomet/36.1-2.75

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences*, *13*(3), 273–304. https://doi.org/10.1207/s15327809jls1303_2

Sadler, T. D. (2004). Informal reasoning regarding socio-scientific issues: A critical review of research. *Journal of Research in Science Teaching*, *41*(5), 513–536. https://doi.org/10.1002/tea.20009

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. (2019). 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

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

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability* (pp. 3–30). SAGE Publications.

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/1177706637

U.S. Global Change Research Program. (2009). *Climate literacy: The essential principles of climate science*. http://downloads.globalchange.gov/Literacy/climate_literacy_highres_english.pdf

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (A. Kozulin, Trans.). Harvard University Press.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *American Statistician*, *73*(S1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

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. https://doi.org/10.2307/20650284

Zangori, L., Peel, A., Kinslow, A., Friedrichsen, P., & Sadler, T. D. (2017). Student development of model-based reasoning about carbon cycling and climate change in a socio-scientific issues unit. *Journal of Research in Science Teaching*, *54*(10), 1249–1273. https://doi.org/10.1002/tea.21404