ASTRONOMY EDUCATION JOURNAL

Astronomy Education Journal, 2022, Vol. 02, No. 1

doi: 10.32374/AEJ.2022.2.1.032ra
ISSN: 2004-2981 · Published in Sweden

RESEARCH ARTICLE

Students' Scientific Evaluations of Astronomical Origins

Archana Dobaria[1] *, Janelle M. Bailey[1], Timothy G. Klavon[2] † and Doug Lombardi[2]

[1]Temple University, Philadelphia, PA, USA ; [2]University of Maryland, College Park, MD, USA

*archana.dobaria@temple.edu
†Now at Black Hills State University

Abstract

Students often encounter alternative explanations about astronomical phenomena. However, inconsistent with astronomers' practices, students may not be scientific, critical, and evaluative when comparing alternatives. Instructional scaffolds, such as the Model-Evidence Link (MEL) diagram, where students weigh connections between lines of evidence and alternative explanations, may help facilitate students' scientific evaluation and deepen their learning about astronomy. Our research team has developed two forms of the MEL: (a) the preconstructed MEL (pcMEL), where students are given four lines of evidence and two alternative explanatory models about the formation of Earth's Moon and (b) the build-a-MEL (baMEL), where students construct their own diagrams by choosing four lines scientific evidence out of eight choices and two alternative explanatory model out of three choices, about the origins of the Universe. The present study compared the more autonomy-supportive baMEL to the less autonomy-supportive pcMEL and found that both scaffolds shifted high school student and preservice teacher participants' plausibility judgments toward a more scientific stance and increased their knowledge about the topics. Additional analyses revealed that the baMEL resulted in deeper evaluations and had stronger relations between levels of evaluation and post-instructional plausibility judgements and knowledge compared to the pcMEL. This present study, focused on astronomical topics, supports our team's earlier research that scaffolds such as the MELs in combination with more autonomy-supportive classrooms may be one way to deepen students' scientific thinking and increase their knowledge of complex scientific phenomena.

Keywords: Astronomy instruction; Astronomy education; Instructional scaffolding; Critical evaluation

1 Introduction

Teaching astronomy topics can be difficult due to spread of misinformation, the presence of alternative conceptions and students' lack of scientific background. Astronomy topics are often abstract and complex, requiring explicit scaffolding of sub concepts through both formal (e.g., activities) and informal (e.g., group discussion) mechanisms. Students who do not have the opportunity to engage with such scaffolding may be more likely to accept alternative conceptual explanations about astronomy phenomena (Lelliott and Rollnick, 2010). Many research studies have suggested a hands-on approach with group discussions helps students gain scientific conceptual understanding (e.g., Plummer and Maynard, 2014). Critique and evaluation play an important role in constructing understanding (Lombardi et al., 2018a).

Students need to deepen their ability to critically evaluate scientific knowledge and weigh alternate explanations (Ford, 2015; National Research Council, 2012). We have been developing instructional scaffolds, called Model-Evidence Link (MEL)

diagrams, to facilitate critical evaluation about Earth, environmental, and space science topics. Formation of Earth's Moon and the origins of the Universe are two topics within space science where students might have difficulty evaluating connections between lines of evidence and alternative explanations. With the Moon formation phenomenon, we developed a MEL around two explanatory alternatives: giant impact theory, which has become the accepted scientific model for Earth's Moon, and capture theory, which does not apply to Earth but likely does for other moons in the solar system. A second complex concept in astronomy is the origin and evolution of the Universe. The Big Bang theory is the accepted scientific model that describes the Universe at the earliest time that we have been able to measure (Coble et al., 2015; Friedman, 1922). Alternative explanations of how the Universe came to be include the steady state model (an early competitor to the Big Bang; (Bondi and Gold, 1948; Hoyle, 1948)) and an explosion of a ball of matter a finite time ago (a common alternative conception; (Bailey et al., 2012; Hansson and Redfors, 2006; Prather et al., 2002)). The aim of each activity is to provide students with detailed evidence, have them consider how that evidence connects to the competing explanations, and then ask students to critically evaluate and make a plausibility judgment about each explanation of the phenomenon.

2 Background Literature

The present study takes inspiration from previous research by Lombardi and colleagues (notably (Bailey et al., 2022; Lombardi et al., 2018a, 2013b; Medrano et al., 2020)) in order to build upon the connection between three fundamental concepts: critical evaluation, plausibility judgment, and conceptual agency. The Plausibility Judgements in Conceptual Change (PJCC) theoretical model posits that critical evaluations about the connections between lines of science evidence and alternative explanations can facilitate more scientific plausibility appraisals and deeper learning (Lombardi et al., 2016c). Although much empirical research has supported the PJCC, most of the topics examined have been related to Earth and environmental science. In this study, we are investigating the efficacy of the PJCC within the context of astronomy instruction and are specifically examining students' levels of evaluation about evidence to explanatory model connections, plausibility judgments about these explanatory modes, and their knowledge of astronomy topics when engaging in scaffolded instruction. The following sections provide a detailed dive into each of these fundamental components of the present study.

2.1 Critical Evaluation

Critical evaluation, in which students look at all possible sides to determine which concept is most plausible, is fundamental to the scientific process (Ford, 2015; National Research Council, 2012; NGSS Lead States, 2013). This process involves students asking questions about how the data or the evidence relates to a given model or explanation. Recent science education reform efforts place this process of critical evaluation at the core of scientific activities that students should engage in during science instruction (National Research Council, 2012, p.45). Students who engage in the critical evaluation process must be reflective about the process of knowledge creation. They must recognize that scientific knowledge derives from collective argumentation, which is a constructive and social process in which people compare, criticize, and revise ideas (Mason et al., 2011; Nussbaum, 2011).

One way to promote evaluative processes is to explicitly engage students in epistemic judgments about knowledge and knowing (Lombardi et al., 2016b). The studies related to the MEL project have focused specifically on the connection between students' evaluation and plausibility judgments of alternative and scientific models for different abstract and/or complex topics. We measure students' evaluation based on how students describe the connections between evidence and alternative models. Previous research has shown that students' evaluations differ qualitatively and that these differences are predictive of post-instructional knowledge (Lombardi et al., 2016a).

Qualitative differences reflect the scientific accuracy of students' evaluation and the reasoning quality of their explanations. Therefore, helping students to become more critically evaluative as they learn about science will potentially lead them to be more scientifically literate. Critical evaluation is stimulated by argumentative discourse activities, where students challenge each other's thinking through questions about the strength of evidence and explanation of those connections (Chin and Osborne, 2010). Because students may not naturally be critically reflective when engaging in collaborative argument, they may need instructional scaffolds to evaluate the quality of explanations (Nussbaum and Edwards, 2011). A promising scaffold that may help students develop deeper levels of evaluative thinking is the MEL diagram, which assists students in effectively coordinating evidence with scientific explanations (Chinn and Buckland, 2012; Lombardi et al., 2013a). The MEL facilitates evaluation and helps students differentiate between evidence and scientific explanations—a scientific reasoning skill with which students often have difficulty (Kuhn and Pearsall, 2000).

2.2 Plausibility Judgments

Plausibility is one of four epistemic judgments that students make about scientific information (Dole and Sinatra, 1998). Plausibility is an epistemic judgment about explanations that is often formed through automatic cognitive processes that facilitate the construction and reconstruction of knowledge both in science and in science classrooms (Dole and Sinatra, 1998; Ceyhan et al., 2021; Lombardi et al., 2013b; Medrano et al., 2020). Lombardi et al. (2013b) described a "plausibility gap" between what students and scientists find plausible for complex socio-scientific issues. Research has found plausibility gaps—such as students not finding human induced climate change as plausible as scientists find it—among middle and high school students (Lombardi et al., 2013b, 2018b,a), undergraduate students (Lombardi and Sinatra, 2010), and elementary and secondary science teachers (Lombardi and Sinatra, 2013). Our versions of the MEL diagram are designed to address plausibility gaps by helping students reappraise their initial judgments via more critical and purposeful evaluations of the connections between lines of evidence and explanations, in light of alternatives (Bailey et al., 2020; Lombardi, 2016). Knowing why a model or a hypothesis is plausible or implausible demonstrates that students have a deeper understanding of a scientific concept (Larrain et al., 2017; Lombardi et al., 2016b). Being able to show how a piece of evidence is linked to a model and give a clear, detailed verbal or written explanation may increase self-efficacy, motivation, and productive attitudes toward learning (Arthurs and Templeton, 2009; Berg, 2014; Brewe et al., 2009; Roemmele, 2017). The MEL project suggests that in simulating scientific strategies, students may develop a deeper understanding of scientific practices and develop critical and analytical thinking and reasoning skills.

2.3 Conceptual Agency

To facilitate agency and greater transfer of critical evaluation skills, we have created an additional graphical scaffold that is complementary to the existing Model-Evidence Link (MEL) diagrams. We designed these scaffolds, "build-a-MELs" (baMELs) to activate conceptual agency, where students are the authors of their own knowledge, accountable to their learning community, and have the authority to make reasoned decisions (Nussbaum and Asterhan, 2016). Based on Patall and colleagues' (2019) suggestions that adolescents' (middle and high school students') classroom engagement is increased when they are allowed more autonomy in instructional settings, we designed the baMELs to be more autonomy-supportive than the preconstructed versions (pcMELs). Increased engagement, particularly around topics that students find interesting and compelling, such as universal origins, may further deepen students' agency as constructors of knowledge (i.e., their conceptual agency; (Reeve and Shin, 2020)).

3 The Present Study

The purpose of the present study was to examine two different forms of the MEL scaffold: the Moon Formation pcMEL, which is less autonomy-supportive, and the Origins of the Universe baMEL, which is more autonomy-supportive. In the Moon pcMEL, students are presented with four lines of scientific evidence and two alternative explanatory models about the formation of Earth's Moon. In the Origins baMEL, students select four lines of scientific evidence from eight possible choices and two alternative explanatory models from three choices about how the Universe originated. Both activities cover socio-scientific topics that align with several disciplinary core ideas, scientific practices, and crosscutting concepts identified in recent U.S. science education reform efforts (National Research Council, 2012). For example, both scaffolds align with the disciplinary core ideas about the nature of the cosmos, which says that "patterns of the apparent motion of the Sun, the Moon, and stars in the sky can be observed, described, predicted, and explained with models. The universe began with a period of extreme and rapid expansion known as the Big Bang. Earth and its solar system are part of the Milky Way galaxy, which is one of many galaxies in the universe" (National Research Council, 2012, p.174).

The primary goal of the present study was to test the impact of both types of MEL scaffolds in (a) promoting students' evaluations when gauging the connections between lines of scientific evidence models and alternative explanations; (b) promoting plausibility appraisals toward a more scientific stance; and (c) deepening students' knowledge of the phenomenon. We were guided by three research questions:

1. What are the levels of students' evaluations when they engage in two instructional scaffolds (i.e., the pcMEL and baMEL) on astronomy topics?

2. How do students' plausibility judgements and knowledge about astronomy topics change over the course of these two instructional scaffolds?

3. How do relations between students' evaluations, plausibility judgements, and knowledge compare between the astronomy pcMEL and baMEL?

Based on theoretical and empirical studies in educational and developmental psychology (e.g., Lombardi et al., 2016c; Patall et al., 2019; Reeve and Shin, 2020) and science and discipline-based educational research (LaDue et al., 2021), we hypothesized the baMEL activity would show more scientific evaluations and plausibility judgements, as well as deeper knowledge, than the pcMEL.

4 Methods

4.1 Participants and Setting

We had two groups of participants (N = 42): (a) high school students (n = 14) and (b) preservice teachers (PSTs; n = 28), both from the mid-Atlantic region in the U.S. The high school student group was enrolled in an Earth science class, taught by an inservice teacher who participated in a summer workshop as described briefly in section 4.3. The preservice teacher group was enrolled in programs for secondary science majors that included teacher preparation and certification eligibility and completed the activities in the first of two required science teaching methods courses taught by a project team member. Although we did not collect individual demographic characteristics for the participants, we did note that each group reflected the demographics of their institutions. High school participants were from a suburban setting near a large city, and reflected a diverse population (White, 56%; Black, 21%; Hispanic of any origin, 14%; multiracial, 6%; Asian or American indigenous, 3%). PSTs were from a university in the large city near the high school (White, 54%; Black, 12%; Asian, 12%; Hispanic of any origin, 7%; American indigenous, International, or unknown, 15%).

4.2 Materials and Design

We implemented the MEL scaffolds in these two different classroom settings, with the pcMEL covering the topic of the formation of Earth's Moon and the baMEL covering the topic of the origin of the Universe. These topics included multiple lines of scientific evidence and explanatory models about space science-related phenomena that students could evaluate. Both topics were part of the curricular scope and sequence in the classrooms that participated in the present study, with all materials developed and validated by experts in science education, master teachers, and astronomers. The MEL and baMEL materials are available online (`https://serc.carleton.edu/mel/teaching_resources/index.html`) for teachers to use. This includes all the materials described below except the Knowledge Surveys (addressed in section 4.2.5), as well as similar activities in other Earth science topics.

4.2.1 Moon Formation pcMEL

Students were introduced to the Moon Formation pcMEL (Figure 1; (Bailey et al., 2016)), which shows two explanatory models saying that (a) the Moon formed after a large object collided with Earth and material from both combined to create the Moon (the scientific explanation, indicated as Model B in the scaffold) and (b) the Moon was an object that came from elsewhere in the solar system and was captured by Earth's gravity (the alternative explanation, indicated as Model A in the scaffold). Note that by "scientific" we mean an explanation that is currently accepted by the broad scientific community; alternative explanations, in general, may have been accepted or strongly considered in the past (e.g., Moon Formation model A, Origins alternative B) or may be explanations that are common misunderstandings (e.g., Origins alternative C). Participants were not told prior to the activity which model (i.e., Model A, or B) is the scientific explanation. The pcMEL also displays four lines of scientific evidence related to the phenomenon, discussing (a) the density of Earth compared to the Moon, (b) collision simulations of other star systems, (c) the inclination of the Moon's orbit compared to Earth, and (d)

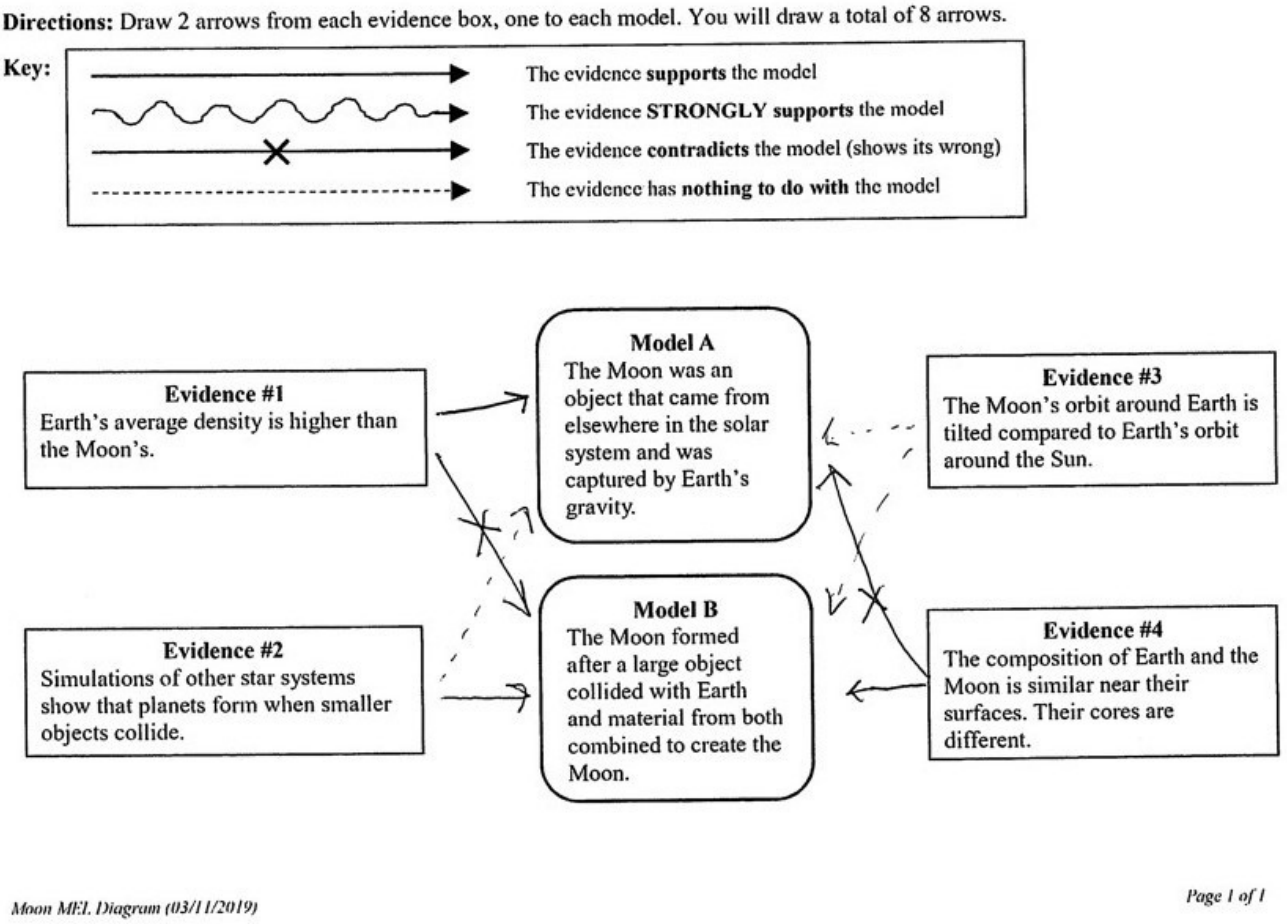


Figure 1. Student's example of the preconstructed Model-Evidence Link (pcMEL) scaffold

structural composition of both Earth and the Moon (Figure 1). Although each line of evidence is presented as one or two sentences on the scaffold, participants were also provided one-page expository texts, with figures and data, elaborating on each line of evidence.

Participants were then instructed to draw different types of arrows from each evidence text to both models based on how well they thought each line of evidence supported each explanatory 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 1).

4.2.2 Origins of the Universe baMEL

In a later lesson, participants examined the beginning and evolution of the Universe with the Origins baMEL (Figure 2; (Bailey et al., 2020)). For this baMEL scaffold, we designed three plausible explanatory models, the scientific explanation that the (a) Universe began via the Big Bang (space, time, and matter came into existence a finite time ago in a hot dense state and has been expanding and cooling ever since; Model A), and two alternative explanations saying that the (b) Universe has always existed in a steady state (Model B) and (c) Universe began as a small ball of matter exploding into a vast empty space (Model C). The eight lines of scientific evidence covered descriptions of the Uniformitarian Principle, simulations of element formation compared to the current composition of the Universe, observed distribution of galaxies, cosmic microwave background radiation, temperature measurements of the Universe, the redshift phenomenon, observations of the Universe's expansion rate, and cosmic temperature profiles.

Participants were first introduced to the lines of evidence and explanatory models before constructing their baMEL. Using one-page expository texts, with figures and data that elaborated on each line of evidence, participants chose two of the explanatory models from the three possible choices, and four lines of evidence from eight possible choices. After indicating their choices on a template, participants completed their self-constructed diagrams by drawing arrows in the same manner as the pcMEL (Figure 2).

4.2.3 Explanation Task: Evaluation Scores

After completing a MEL diagram (either pcMEL or baMEL), participants completed what we call the "Explanation Task" (Figure 3). Participants picked three of the connections that they drew from the MEL activity and wrote explanations about why they drew a particular type of arrow (i.e., an explanation of their evaluation of the strength between a particular line of evidence and a particular model). Using a scoring system and rubric developed by Lombardi et al. (2016a), coders rated explanations for different levels of evaluation: 1 = Erroneous, 2 = Descriptive, 3 = Relational, and 4 = Critical. The categories established well-defined levels of evaluation to represent the accuracy and elaboration 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, with full consensus reached after consultation. The final evaluation score was the average of the consensus scores for each explanation.

103 001 199

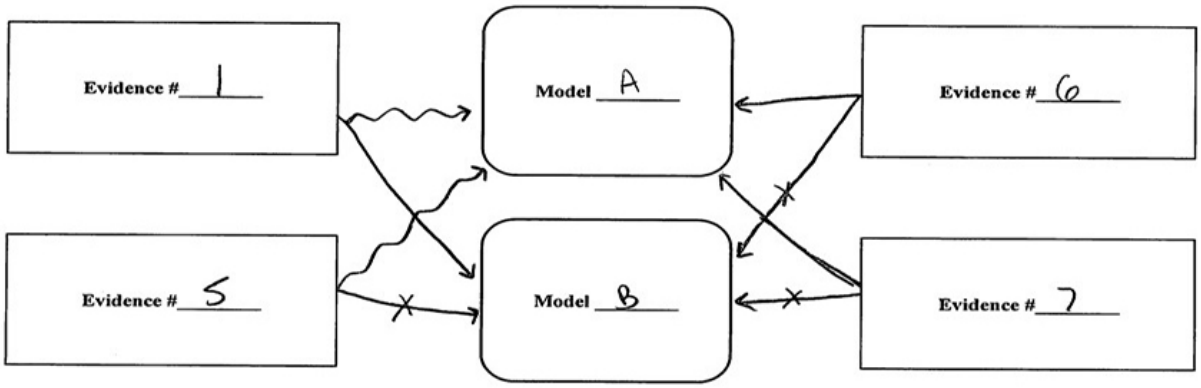
Name: _____ **Date:** _____ **Teacher:** _____ **Period:** _____

If you worked with other students, their name(s): _____

Directions: Write the number of each evidence you are using and for each model you have selected in the boxes below. Then draw 2 arrows from each evidence box, one to each model. You will draw a total of 8 arrows.

Key:

- The evidence **supports** the model
- The evidence **STRONGLY supports** the model
- The evidence **contradicts** the model (shows its wrong)
- The evidence has **nothing to do with** the model



baMEL Worksheet (02/11//2018) *Page 1 of 1*

Figure 2. Student's example of the build-a-MEL (baMEL) scaffold.

Provide a reason for three of the arrows you have drawn. **Write your reasons for the three most interesting or important arrows.**

A. Write the number of the evidence you are writing about.
B. Circle the appropriate word (**strongly supports | supports | contradicts | has nothing to do with**).
C. Write which model you are writing about.
D. Then write your reason.

1. Evidence # 4 **(strongly supports) | supports | contradicts | has nothing to do with Model** B **because:**

Evidence #4 strongly supports model B because if a large object were to collide with earth & the earth material got shot to it & went into orbit of earth then the moon would definitely have mostly the same composition.

2. Evidence # 1 **(strongly supports) | supports | contradicts | has nothing to do with Model** B **because:**

Evidence #1 supports model B because if something collides with earth to create the moon then the center of the moon would be close to the same density as earth but more rocks collided with the moon making the less dense crust of the moon

3. Evidence # 2 **strongly supports | (supports) | contradicts | has nothing to do with Model** A **because:**

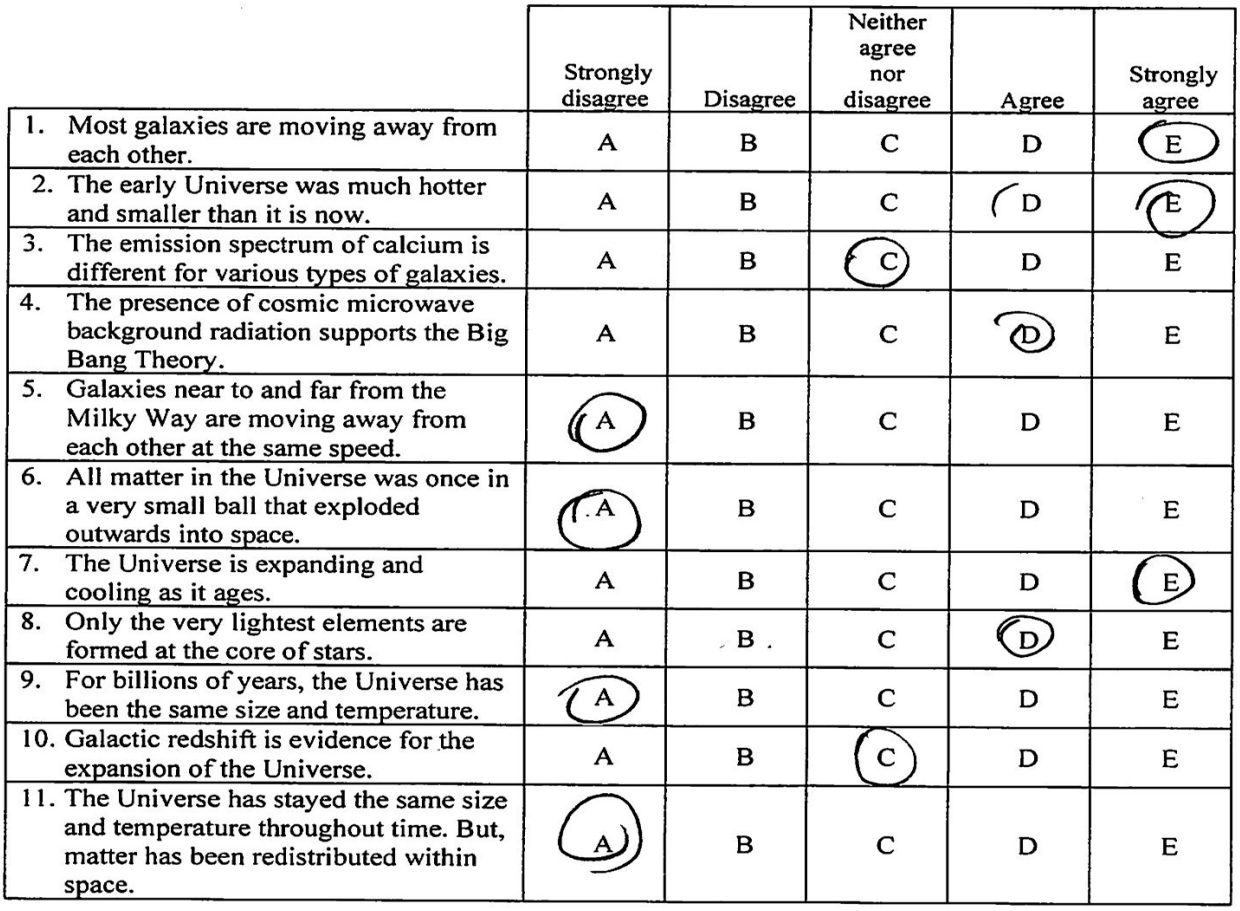
Model A states the rocks in the early forming Solar system collided with each other to make the moon & that would make sense with evidence #2

Circle the plausibility of each model. [Make two circles, one for each model.]

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

Moon MEL Explanation Task (08/02/2015) *Page 1 of 1*

Figure 3. Student's example of a completed explanation task.



Below are statements about the origins and evolution of the Universe. Rate the degree to which you think that *astronomers* agree with these statements.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1. Most galaxies are moving away from each other.	A	B	C	D	E
2. The early Universe was much hotter and smaller than it is now.	A	B	C	D	E
3. The emission spectrum of calcium is different for various types of galaxies.	A	B	C	D	E
4. The presence of cosmic microwave background radiation supports the Big Bang Theory.	A	B	C	D	E
5. Galaxies near to and far from the Milky Way are moving away from each other at the same speed.	A	B	C	D	E
6. All matter in the Universe was once in a very small ball that exploded outwards into space.	A	B	C	D	E
7. The Universe is expanding and cooling as it ages.	A	B	C	D	E
8. Only the very lightest elements are formed at the core of stars.	A	B	C	D	E
9. For billions of years, the Universe has been the same size and temperature.	A	B	C	D	E
10. Galactic redshift is evidence for the expansion of the Universe.	A	B	C	D	E
11. The Universe has stayed the same size and temperature throughout time. But, matter has been redistributed within space.	A	B	C	D	E

Figure 4. A completed Origins Knowledge Survey.

4.2.4 Model Plausibility Ratings: Plausibility Judgment Scores

For both the pcMEL and baMEL, participants were instructed to rate the plausibility of all explanatory models both pre- and post-instruction. Pre instruction, students read stand-alone documents that provided each model and short explanations before rating the plausibility of each model in a manner similar to the bottom section on Figure 3. 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. (2013b). For the Origins baMEL, participants recorded their plausibility judgments for all three explanatory models, while for the Moon pcMEL, students recorded their plausibility judgments for the two explanatory models. Because the Moon pcMEL offered only two explanatory models, we calculated scores as the rating of the scientific model minus the alternative. The Origins baMEL offered three different explanatory models (scientific and two alternative explanations), and therefore, we calculated three different scores: scientific minus alternative explanation 1 (i.e., Model A - Model B); scientific minus alternative explanation 2 (i.e., Model A - Model C); and alternative explanation 2 minus alternative explanation 1 (i.e., Model C - Model B). 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 (or alternative 2 more plausible than alternative 1), with negative scores indicating participants judged the alternative explanation as being more plausible than the scientific (or alternative 1 more plausible than alternative 2).

4.2.5 Knowledge Survey: Knowledge Scores

Participants completed a multi-item knowledge survey instrument (at pre- and post-instruction). The Origins Knowledge Survey (Figure 4) contained 11 items and the Moon Formation Knowledge Survey (Figure 5) contained 8 items. 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, Sinatra et al. (2013b). At least one question in each set addressed each line of scientific evidence. Questions were constructed in two different formats: Some statements were negatively worded (i.e., in effect scientists would disagree with these knowledge statements) and we reverse coded these statements. McDonald's omega coefficients (ω)—a measure of reliability that that does not assume equal factor loadings (as does Cronbach's alpha) and therefore, is a more generalized estimator (Hayes and Coutts, 2020)—were used to examine if the knowledge scale for each scaffold and time points were sufficiently reliable (Moon pre: $\omega = 0.67$, post: $\omega = 0.72$; Origins pre: $\omega = 0.63$, post: $\omega = 0.66$). The interpretation of these values is typically similar to alpha; here the reliability of the knowledge survey would be considered acceptable or fair (George and Mallery, 2009; Tabachnick and Fidell, 2007).

Below are statements about the Moon. Rate the degree to which you think that *planetary scientists* agree with these statements.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1. Studying moon rocks brought back by Apollo astronauts can tell us about the history of the Moon.	A	B	C	D	E
2. The Moon's orbit around Earth is slightly tilted compared to Earth's orbit around the Sun.	A	B	C	D	E
3. Earth's density is lowest in the core and increases as you go out toward the crust.	A	B	C	D	E
4. Craters on the Moon and Mercury show that these objects have been impacted by many smaller objects over billions of years.	A	B	C	D	E
5. Scientists have created good models about how the Moon formed.	A	B	C	D	E
6. The Moon's core contains about the same percentage of iron as Earth's core.	A	B	C	D	E
7. Planets and moons are formed through many collisions of smaller objects.	A	B	C	D	E
8. Seismic measurements by Apollo astronauts show that there are three main layers in the Moon.	A	B	C	D	E

Figure 5. A completed Moon Formation Knowledge Survey.

4.3 Data Collection and Procedures

During the summer prior to the present study, inservice classroom teachers participated in a three-day professional development workshop with the project team; one of these teachers later allowed us to collect data in their classroom with their high school students (i.e., the high school participants in the present study). The workshops focused on introducing and practicing using the pcMEL and baMEL activities, going over the pedagogical aspects of the MEL activities for effective classroom implementation, and planning for the upcoming year's implementation. To maintain some uniformity in instruction, the teachers agreed to introduce each strategy introduced at the workshop for effective classroom implementation. Teachers also agreed to follow the lesson plans as specified at the workshop and as found in the teacher guide. Participating high school students completed all the activities over the course of an instructional unit focused on astronomy topics. Participating preservice teachers engaged in the same activities during their science teaching methods class.

Prior to either treatment, participants completed the Plausibility Ranking Task (PRT) as an introduction to the ideas of scientific evaluations and plausibility judgments. This task asked participants to rank the four different evaluation categories (evidence strongly supports an explanatory model, evidence supports an explanatory model, evidence has nothing to do with an explanatory model, and evidence contradicts an explanatory model) on their importance in making judgment of an explanation's plausibility. These four categories are the same as the arrows that participants would later draw on their MEL diagrams. After ranking the importance of each form of evidence, participants read a small passage on falsifiability positing that scientific explanations cannot be proven but are rather disproven through opposing evidence. Participants then re-ranked the four evaluation categories again. This task introduced the idea of scientific evaluations and plausibility judgments, but these ranked data were not used in the present study's analyses. The PRT leads students to understand the strong role of contradictory evidence, and preliminary research in development indicates that many students move from believing "strongly supports" is most important to "contradicts" after this task. Additionally, the identification of contradictory evidence has shown to be related to stronger evaluations (Lombardi et al., 2016a).

Figure 6 shows the procedure for each treatment that takes place after the Plausibility Ranking Task. For a given activity, participants began by completing the knowledge survey and Model Plausibility Ratings (pre) for each explanatory model on the topic. The teacher also engaged students in the class in an unscripted short discussion of the models and the idea of plausibility to clarify misunderstandings and address general questions about the topic. For the pcMEL, participants read the four pieces of evidence and completed the MEL diagram in small groups after discussing the relationships between each evidence and model. After drawing their diagrams, participants completed

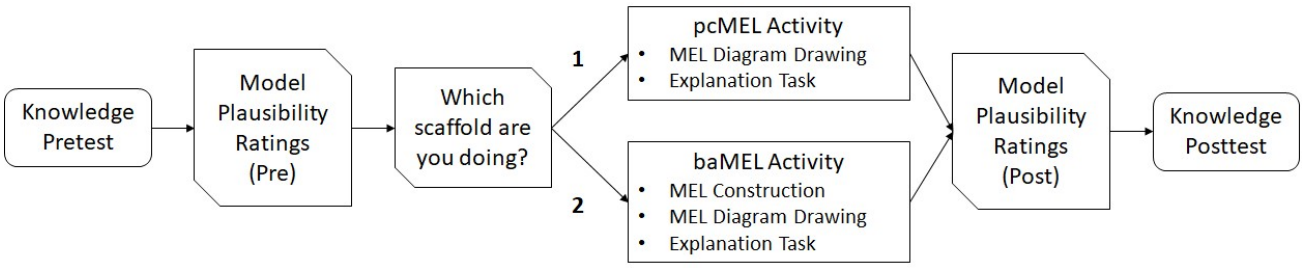


Figure 6. MEL activities in chronological order.

the Explanation Task individually. Participants then re-rated the plausibility of each model. The activity ended with students completing the knowledge survey about the topic (post). For the baMEL, the activity followed a similar path except participants constructed their diagrams before drawing on them. During the construction process, participants first read the eight lines of scientific evidence and choose which four they would like to keep after discussing with the group. Participants also decided which two models, out of a possible three, that they would include on their diagrams.

5 Results

We present our results in three sections. The first section addresses Research Question 1: *What are the levels of students' evaluations when they engage in two instructional scaffolds (i.e., the pcMEL and baMEL) on astronomy topics?* The second section addresses Research Question 2: *How do students' plausibility judgments and knowledge about astronomy topics change over the course of these two instructional scaffolds?* These two sections collectively represent a fine-grained comparative analysis of the effectiveness of the instructional treatments (pcMEL and baMEL). The third section addresses Research Question 3: *How do relations between students' evaluations, plausibility judgments, and knowledge compare between the astronomy pcMEL and baMEL?* We are analyzing the relationships between the variables present in the MEL diagram activities in this third question, a larger-grained analysis than the prior two questions. The analyses used take into consideration the research question and the size of the data set. In each case we provide effect sizes and discuss the meaning of that effect size in practical terms. Research Questions 1 and 2 are looking at the impact on dependent variables over time. For this we calculated effect size using partial eta-squared (η_p^2), which reflects the percentage of the variance in the dependent variable explained by the independent variables in a sample. Research Question 3 is looking at the relation between dependent variables. Here we used Cohen's f-squared measurements, an effect size often used for simple and multiple linear regression, to gauge the relative strength of relational pathways. In calculating Cohen's f-squared, we employed a slightly different technique, as computed within the Warp PLS 7.0 software, than that proposed by Cohen (1988) to ensure that variable weighting did not result in effect size bias (Kock, 2020). We gauged the robustness (i.e., strength) of our results using effect size indices when differences and/or relations were statistically significant. For example, we calculated effect sizes (η_p^2) for the analyses of variances (ANOVAs), and used .01, .09 and .25 to gauge small, medium, and large effect sizes, respectively (Cohen, 1988). These "rules of thumb" are likely conservative (i.e., too restrictive) based on a recent review of educational research conducted by Kraft (2020). Because of the relatively small sample size, we ran both ordinary least squares (ANOVAs) and categorical analyses; however, there was no meaningful difference in outcomes, and therefore, the following discussion includes only ANOVA results. We also screened the data and found that they met the assumptions inherent within ordinary least squares tests (i.e., normality, linearity, and homogeneity of variance).

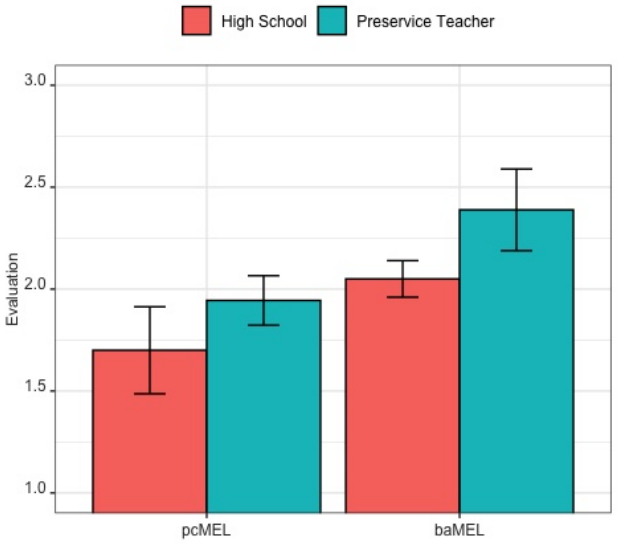


Figure 7. Evaluation scores by scaffold type and level. Evaluation scores ranged from 1 = erroneous description to 4 = critical evaluation for each instructional treatment. Error bars indicate ±1 standard error.

5.1 Research Question 1: Evaluation by Scaffold

We conducted a repeated measures ANOVA with evaluation scores as the dependent variable, scaffold type (pcMEL and baMEL) as the within-subjects variable, and level (high school and undergraduate preservice teacher) as a between-subjects factors (Figure 7). The ANOVA indicated that there was not a statistically significant interaction between scaffold and level ($p = .791$); however, that the main effect of scaffold revealed a medium effect size, with the baMEL evaluation scores ($M = 2.27$, $SD = 0.71$) significantly greater than pcMEL scores ($M = 1.86$, $SD = 0.58$), $F(1,26) = 5.06$, $p < .033$, $\eta_p^2 = .163$.

5.2 Research Question 2: Changes Over Time

5.2.1 Plausibility Judgment Scores

We conducted a repeated-measures ANOVA for the pcMEL, with plausibility scores (scientific explanation minus alternative explanation) as the dependent variable, time as the within-subjects factor, and level as the between-subjects factor (Figure 8). The interaction between time and level revealed a medium effect size, $F(1,26) = 3.41$, $p = .073$, $\eta_p^2 = .116$, with the relatively small sample size probably resulting in a p-value typically con-

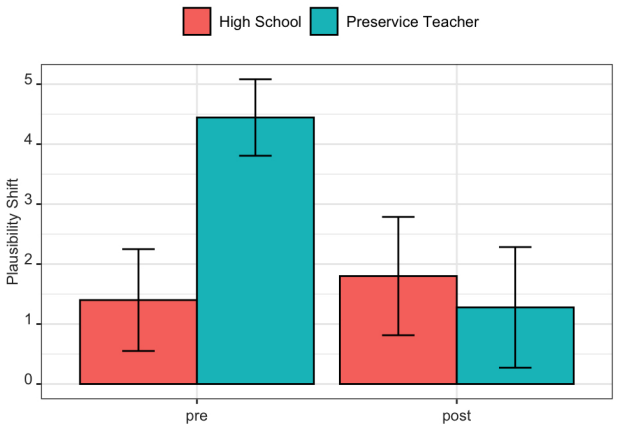


Figure 8. Plausibility scores by level for the pcMEL. Plausibility scores could range from -9 (highly implausible) to +9 (highly plausible). Error bars indicate ±1 standard error.

sidered above the statistically significant threshold. However, because of the robustness of the effect size, we conducted a follow up simple effects analysis, which revealed that there was a statistically significant increase in preservice teacher participants' plausibility scores from pre- (M = 1.28, SD = 4.29) to post-instruction (M = 4.44, SD = 2.71), $F(1,26)$ = 6.13, p = .024, η_p^2 = .218 (medium effect size). However, there was no statistically significant change in high school student participant scores from pre- to post-instruction (p = .798).

We next ran three repeated measures ANOVAs for the baMEL with plausibility scores (i: scientific explanation, Model A, minus the steady-state explanation, Model B; ii: scientific explanation, Model A, minus the explosion explanation, Model C; and iii: the explosion explanation, Model C, minus the steady-state explanation, Model B) as dependent variables, time as within-subjects factor, and level as between-subjects factor (Figure 9). For the first ANOVA (Model A - Model B), the interaction was not significant (p = .620), but there was a main effect for time (large effect size), with a significant increase in plausibility scores from pre- (M = 3.89, SD = 2.41) to post-instruction (M = 5.64, SD = 2.33), $F(1,26)$ = 21.7, p < .001, η_p^2 = .455. For the second ANOVA (Model A - Model C), the interaction was also not significant (p = .731), but again there was a main effect for time (large effect size), with a significant increase in plausibility scores from pre- (M = 0.50, SD = 1.86) to post-instruction (M = 2.11, SD = 2.25), $F(1,26)$ = 11.3, p = .002, η_p^2 = .303. For the third ANOVA (Model C - Model B), neither the interaction nor the main effect for time was significant (all p-values > .428)

5.2.2 Knowledge Scores

We conducted two repeated-measures ANOVA for the pcMEL and baMEL, with knowledge score as the dependent variable, time as the within-subjects factor, and level as the between-subjects factor (Figure 10). For the pcMEL, the interaction between time and level was not significant (p = .202), but there was a main effect for time (large effect size), with a significant increase in knowledge scores from pre- (M = 3.72, SD = 0.42) to post-instruction (M = 4.02, SD = 0.45), $F(1,26)$ = 14.2, p < .001, η_p^2 = .354. For the baMEL, the interaction was also not significant (p = .313), but there was also a main effect for time (large effect size), with a significant increase in knowledge scores from pre- (M = 3.32, SD = 0.29) to post-instruction (M = 3.66, SD = 0.40), $F(1,26)$ = 19.9, p < .001, η_p^2 = .433.

5.2.3 Summary for Research Question 2: Changes Over Time

Both the pcMEL and baMEL resulted in pre- to post-instructional shifts in plausibility judgments toward the scientific and increases in knowledge, in almost all cases. The exception was that high school students did not show a significant shift in plausibility judgments for the pcMEL. Other than that, there was no meaningful difference between levels (high school and preservice teachers).

5.3 Research Question 3: Relations Between Variables

We used a multi-faceted approach when analyzing Research Question 3: *How do relations between students' evaluations, plausibility judgements, and knowledge compare between the astronomy pcMEL and baMEL?* Structural equation modeling examines relations between variable pathways. Outputs of the analysis include beta weights, which indicate the strengths of relational pathways, as well as other indicators to help gauge how well our hypothesized model of these relations aligns with the data. We used WarpPLS 7.0 (Kock, 2020) for these analyses, which is a partial least squares structural equation modeling (PLS-SEM) tool that uses ranked data and is distribution free (Lombardi et al., 2016b). This reduces standard error and increases statistical power for smaller sample sizes such as our (Kock, 2020). Although the use of PLS-SEM has been criticized in the past (Goodhue et al., 2012), Kock (2020) suggested 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. Therefore, to increase validity of the results, we also employed jackknifing as the resampling technique for PLS-SEM. Jackknifing is a process that reduces standard error and may increase statistical power by removing one or more indicators at a time and replacing them with partial estimates (Abdi and Williams, 2010; Quenouille, 1949; Tukey, 1958). This replacement technique sought to increase the predictive ability of the PLS-SEM (Kock, 2020). We made model

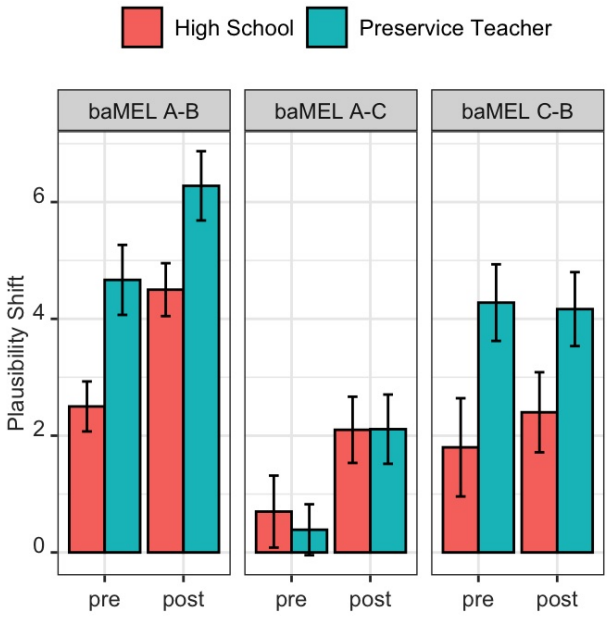


Figure 9. Plausibility scores by level for the baMEL. Plausibility scores could range from -9 (highly implausible) to +9 (highly plausible). Error bars indicate ±1 standard error

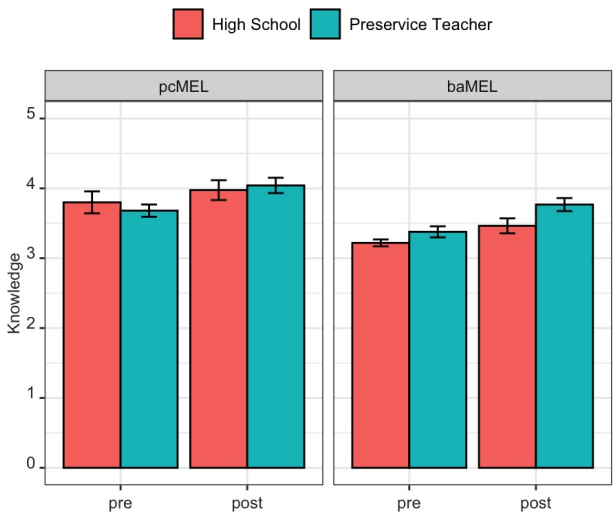


Figure 10. Knowledge scores by level for two scaffold types. Knowledge scores could range from 1 to 5. Error bars indicate ±1 standard error.

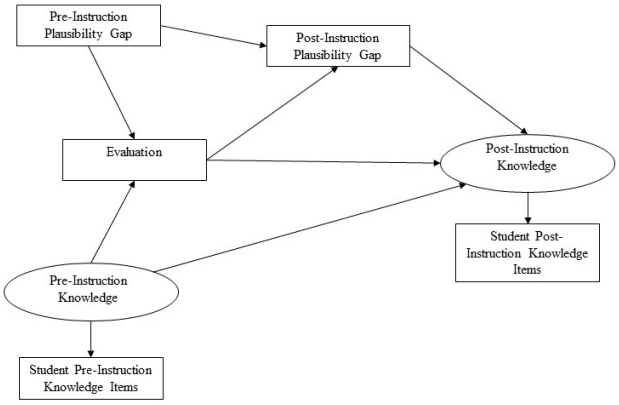


Figure 11. Initial PLS-structural equation model relating plausibility, evaluation, and knowledge. Indicators (i.e., observed values) are designated by rectangles and constructs (i.e., derived values) are designated by ovals.

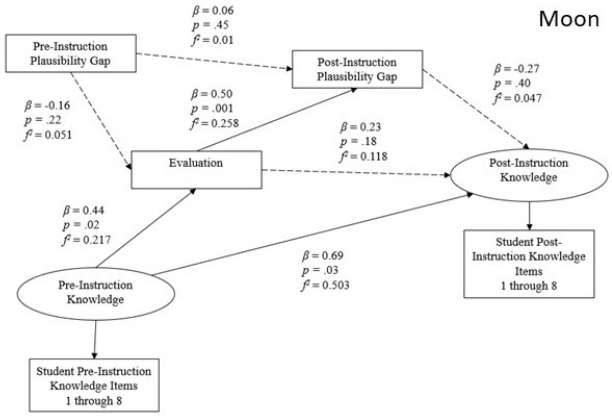


Figure 12. PLS-structural equation model relating the pcMEL evaluation, plausibility, and knowledge scores. Indicators (i.e., observed values) are designated by rectangles and constructs (i.e., derived values) are designated by ovals.

comparisons using Tenenhaus Goodness of Fit (GoF), which suggests how well different subsets of the data can be explained by the model (Henseler and Sarstedt, 2013). Our decision to employ these analytical techniques (i.e, PLS-SEM with jackknifing, Tenenhaus Goodness of Fit) was to allow us to make more generalized assumptions about how students used these instruments to make scientific evaluations, judge model plausibility, and construct science knowledge.

After completing the PLS-SEM, we implemented a holistic approach to evaluating the relationships formed by the model, considering 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, 2020)). Wasserstein et al. (2019) implored us to not "believe that an association or effect is absent just because it was not statistically significant" (p. 1). This more holistic approach provided 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.

5.3.1 Construction of PLS-SEM

We constructed models for each topic, the Moon Formation pcMEL and the Origins baMEL (Figure11). WarpPLS constructed the students' pre- and post-instruction knowledge scores using the individual knowledge survey items as indicators. We ran the model using the jackknifing resampling technique to account for the small sample size (N = 42).

5.3.2 Pre-Constructed MEL (pcMEL) PLS-SEM

We found that the pcMEL PLS-SEM produced a good fit (Tenenhaus GoF = 0.510, large effect size ≥ 0.36). The model produced strong relationships from pre-instruction knowledge (PrK) to evaluation (E) and to post-instruction knowledge (PoK), as well as a very strong relationship between evaluation and post-instruction plausibility (PoP) (Table 1). Holistically, the values of the other links in the initial model indicated that we could remove them. Their overall values of strength, importance, and significance informed our decision to exclude them from the model (Figure 12).

The importance of both the PrK-E link (β = 0.44, f^2 = 0.217, p = .02) and the PrK-PoK link (β = 0.69, f^2 = 0.503, p = .03) is not surprising. Previous projects (e.g., Braasch and Goldman, 2010; Klosterman and Sadler, 2009) have noted the role of prior knowledge to post-instruction knowledge gains, particularly those with the use of analogy, as analogies provide a frame of reference for developing mental models about a scientific phenomenon (Norman, 1983). The E-PoP link was particularly strong (β = 0.50, f^2 = 0.258, p = .001), which is supported by past research (Lombardi et al., 2018a,b; Medrano et al., 2020), suggesting that the students' evaluative efforts may have a strong impact on their plausibility reappraisal.

Our analysis meant the exclusion of the direct link between post-instruction plausibility and post-instruction knowledge because there was no justification to support keeping it in the model. However, past research featuring the MELs, but focusing on other topics (Lombardi et al., 2018a,b; Medrano et al., 2020), suggested that such a strong relationship between evaluation and plausibility gains may also have an impact on knowledge gains, which for the present study was not evident in this small sample.

5.3.3 Build-a-MEL (baMEL) PLS-SEM

The baMEL PLS-SEM also produced a large Tenenhaus Goodness of Fit value (GoF = 0.462), suggesting that the model fit well with the observations. The relevant links that remained in the model were pre-instruction plausibility (PrP) to post-instruction plausi-

Table 1. PLS-structural equation modeling β weights, effect sizes, and significance values for the 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	-0.16	0.051	.22	-	-	-	-	-	-	0.44	0.217	.02
Post-Instruction Plausibility	0.06	0.01	.45	0.50	0.258	.001	-	-	-	-	-	-
Post-Instruction Knowledge	-	-	-	0.23	0.118	.18	-0.27	0.047	.40	0.50	0.414	.03

Note: β represents β weights, f^2 represents the WarpPLS approximation of Cohen's *f*-squared effect size, and *p* represents *p*-value.

Table 2. PLS-structural equation modeling β weights, effect sizes, and significance values for the 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	0.13	0.226	.28	-	-	-	-	-	-	0.43	0.215	.06
Post-Instruction Plausibility	0.44	0.218	.05	0.20	.062	.09	-	-	-	-	-	-
Post-Instruction Knowledge	-	-	-	0.24	0.076	.15	0.50	0.261	.01	0.05	0.009	.44

Note: β represents β weights, f^2 represents the WarpPLS approximation of Cohen's *f*-squared effect size, and *p* represents *p*-value.

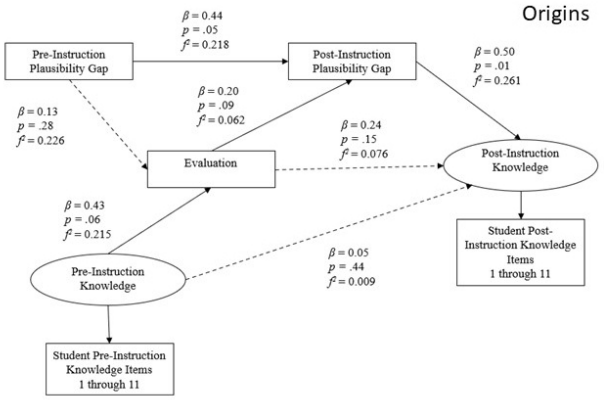


Figure 13. PLS-structural equation model relating the baMEL evaluation, plausibility, and knowledge scores. Indicators (i.e., observed values) are designated by rectangles and constructs (i.e., derived values) are designated by ovals.

bility, pre-instruction knowledge to evaluation, and evaluation to post-instruction plausibility (Table 2). The link between post-instruction plausibility to post-instruction knowledge was particularly powerful, which supports the relationship between post-instruction plausibility and post-instruction knowledge found in previous studies (Figure 13; (Lombardi et al., 2018a,b; Medrano et al., 2020)).

The PrK-E link was important (β = 0.43, f^2 = 0.215, *p* = .06) in the Origins PLS-SEM, with a moderate β weight and effect size and a small effect due to randomness. This is not unusual as students' prior knowledge often is a driving factor in their understanding of the analogies that help them use explanatory models (Braasch and Goldman, 2010; Norman, 1983). Both the PrP-PoP link (β = 0.44, f^2 = 0.218, *p* = .05) and E-PoP link (β = 0.20, f^2 = 0.062, *p* = .09) had a moderate impact in post-instruction plausibility. However, the PoP-PoK link was quite impactful (β = 0.50, f^2 = 0.261, *p* = .01), leading us to conclude that plausibility judgments are strongly related to knowledge gains in this activity. Interestingly, the PrK-PoK link was not supported in this model, which may be a strong indicator that participants' pre- to post- instructional shifts in their plausibility judgments (i.e., plausibility reappraisal) might have been a major factor involved in learning via the baMEL.

5.3.4 PLS-SEM Comparison

We used two analyses to compare the two PLS-SEM results. First, we looked at the overall model fit (i.e., Tenenhaus Goodness of Fit; GoF). In this case, each model had a high GoF (> 0.36), which allows us to consider the models to be highly representative of the observations made during the pcMEL and baMEL activities. We also compared the effect sizes of the analogous links from each of the PLS-SEMs (Table 3; (Medrano et al., 2020)).

Table 3. Effect size comparison.

Link Description	Effect Size (f^2)	
	pcMEL	baMEL
Pre-instruction Plausibility—Evaluation	0.51	0.226*
Pre-instruction Plausibility—Post-instruction Plausibility	0.01	0.218*
Pre-instruction Knowledge—Evaluation	0.217*	0.215
Pre-Instruction Knowledge—Post-instruction Knowledge	0.503*	0.009
Evaluation—Post-instruction Plausibility	0.258*	0.062*
Post-instruction Plausibility—Post-instruction Knowledge	0.047	0.261*
Evaluation—Post-instruction Knowledge	0.118	0.076

Note: * Link included in final PLS-SEM

When comparing the structure of the two PLS-SEMs, we found that, though the pcMEL has a stronger relationship between evaluation and post-instruction plausibility, the overall effectiveness of the baMEL was more desirable. This desirability is based upon our hypothesis that the baMEL activity will engage student agency and, thus, provide greater learning opportunities. The PoP-PoK link gave us a strong indication that, in this case, the baMEL provided the students with a better environment to

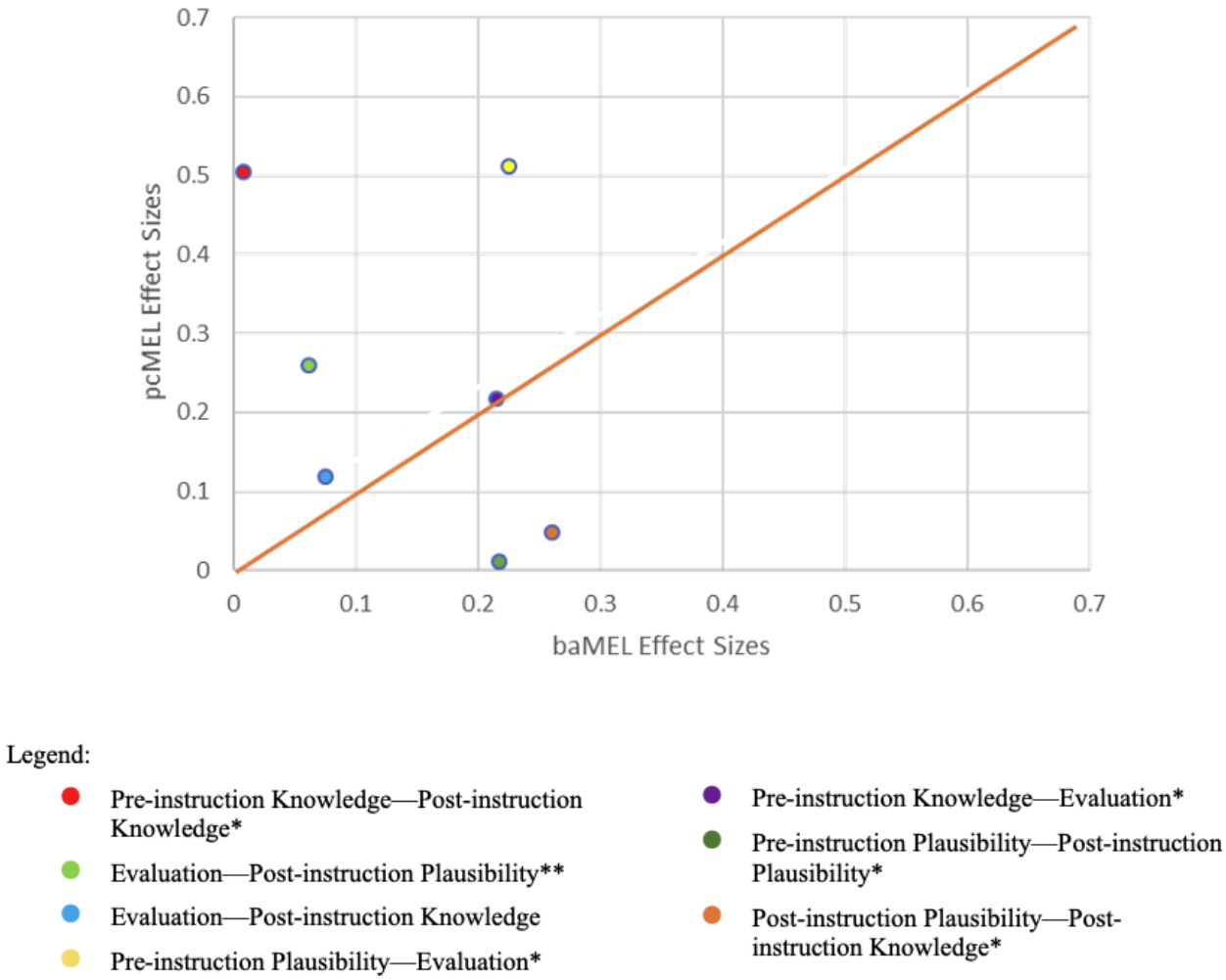


Figure 14. Figure 14: Comparison of effect sizes. *Denotes link included in one PLS-SEM. **Denotes link included in both PLS-SEMs.

use their evaluations to shift their plausibility ratings toward the scientifically accepted model. Figure 14 shows this effect size comparison on a coordinate graph using a reference line with a slope of m = 1. Items above and to the left of the line favor the pcMEL, while items below and to the right of the line favor the baMEL. Whether this positioning on the graph is desirable or not depends on each link item. For example, the positioning of the PoP-PoK point was quite desirable as the post-instruction values favored the baMEL. The items that include pre-instruction plausibility, PrP-E and PrP-PoP, are less important as it is desirable to have student evaluation scores drive their plausibility reappraisal and knowledge gains (Lombardi et al., 2016c).

6 Discussion

Our aim in conducting the present study was to compare two different instructional scaffolds to learn about complex astronomical topics: (a) formation of Earth's Moon and (b) origins of the Universe. The first scaffold, the preconstructed MEL (pcMEL), is less autonomy supportive, with lines of scientific evidence and alternative explanatory models provided to participants. The second scaffold, the build-a-MEL (baMEL), is more autonomy supportive, with participants choosing their lines of scientific evidence and alternative explanatory models from a provided set. Results revealed that both the pcMEL and baMEL resulted in participants shifting their plausibility judgments toward a more scientific stance and deepening their knowledge about each topic, which aligns with previous investigations comparing MELs that cover other topics (i.e., the climate crisis (Bailey et al., 2022); past environments and subsurface processes (Klavon et al., 2021); water resources (Medrano et al., 2020)). The baMEL resulted in more scientific evaluation than the pcMEL, although in both cases participants still had somewhat descriptive evaluations, which tend to superficially connect lines of evidence with explanations (Lombardi et al., 2013b). The effect sizes, which were medium to large in all cases, would suggest that the single activity (either pcMEL or baMEL) that takes only a class period or two can make a difference of about a third of a letter grade (e.g., B- to B). Further, in most cases there were no differences in outcomes between the two different participant levels, high school students and undergraduate preservice teachers. However, structural equation modeling revealed an overall advantage for the baMEL compared to the pcMEL because relations between variables were generally more robust. All in all, we can conclude that within the context of the present study, the baMEL (a more autonomy-supportive instructional scaffold) resulted in participants' greater levels of evaluation in gauging the connections between lines of scientific evidence and alternative explanations, which related to stronger shifts in plausibility toward a more scientific stance and deeper knowledge than the pcMEL (a less autonomy-supportive instructional scaffold).

After engaging in the Origins baMEL, participants shifted their plausibility toward the scientific explanation of universal

origins (the Big Bang, Model A) compared to the alternative, steady state explanation (Model B). However, participants also experienced similar shifts toward the alternative explosion explanation (Model C) when compared to Model B. Further, there were no meaningful plausibility shifts when comparing the plausibility of Model A to Model C, post-instruction. The explanatory difference between Model A and Model C is subtle—the expansion of space itself versus an explosion of matter into existing space, respectively. Because of the apparent similarity of these two models, students may not have fully understood the details in some of the lines of scientific evidence that provide greater support for the Big Bang (Aretz et al., 2016; Cardinot and Fairfield, 2021; Hansson and Redfors, 2006; Prather et al., 2002; Trouille et al., 2013). These results suggest that it is easier to make a distinction between explanations when there is a greater contrast and greater plausibility gap (Lombardi et al., 2016c).

6.1 Limitations

As with all research, including classroom-based research, the present study has a few limitations that call for some caution in interpreting the results. Although our analysis techniques demonstrated sufficient power for making our statistical inferences, with generally moderate effect sizes, the sample size in the present study was relatively small. This is often the case with educational research studies, particularly those involving interventions such as the MELs that require appreciable data collection. Specifically, the results were confined to two classroom settings and generalizing these findings is not warranted. However, our findings are consistent with other research studies using MELs with different topics. This does give us some confidence that our results are valid in many contexts, particularly when learning about complex scientific topics where there may be a large gap between what scientists and learners find plausible (Lombardi et al., 2016c; Sinatra and Lombardi, 2020).

The two scaffolds covered two different astronomical topics (formation of Earth's Moon and origins of the Universe), and we acknowledge that topic difference may have had some influence on the present study's results. By conducting repeated measures analyses examining changes in scores at pre- and post-instruction, as well as interpreting our results via a more holistic approach (e.g., using effect size to gauge strength of results and practical significance), we were largely able to statistically control for topic difference. However, the topic of the Moon's formation may be less controversial than the universal origins, and therefore, the degree of cognitive and implicit commitment toward the Moon may be less than toward the Universe's origin. This lack of commitment and interest may have further lessened participants' engagement, therefore lowering the relation between post-instructional plausibility judgments and Moon knowledge (Lombardi et al., 2021; Sinatra and Lombardi, 2020).

A final potential limitation arises from the implementation of the activities. Based on the Parsons et al. (2017) comprehensive review of educational research literature, the project team supports teachers to have full leeway in making adjustments to the details of the implementation (e.g., adapting lessons to best suit the needs of their students and specific context), rather than meeting external fidelity criteria. Teachers are encouraged to use the professional development training and Teacher's Guide as a strong starting point. Furthermore, the teachers whose students provided data for the present study were involved in the project for multiple years, and this was not their first time implementing either activity (and in fact some had used similar pcMEL or baMEL activities on different topics). Although it is possible that this makes a difference in the research outcomes, it is more important that the classroom learning be maximized and any disruption due to the research minimized.

6.2 Implications

Astronomy is a topic with high interest for both K-12 (e.g., Krstovic et al., 2008) and undergraduate students (Bailey et al., 2017; Fraknoi, 2001), but to tap into this popularity may require new instructional strategies and materials. One of the most challenging tasks that educators have is trying to keep their students cognitively, behaviorally, and emotionally engaged in their schoolwork (Sinatra et al., 2015). In the classroom, being critical involves scientifically evaluating the validity of evidence and explanations through epistemic judgments (e.g., reliability and trustworthiness of evidence and the plausibility of explanations). With regard to how well evidence supports explanations, we suggest that science classrooms should practice purposeful evaluation of these connections, particularly in light of alternative explanations, which students may find more plausible than the scientific, such as the alternative explosion model (Model C) in the Origins of the Universe baMEL (Coble et al., 2015; Prather et al., 2002; Trouille et al., 2013). The present study and our prior research have shown that instructional scaffolding can facilitate narrowing the plausibility gap between scientific and alternative explanations.

Scaffolding that also facilitates evaluation of information sources, such as lateral reading (McGrew, 2020), may also shift students toward more critical evaluations that helps them to better determine and trust the validity of scientific evidence (Sinatra et al., 2015). Lateral reading is a strategy from civics education that involves opening multiple tabs on an internet browser to gather information about the source of a reading in a manner similar to what is used by professional fact-checkers (McGrew, 2020). Thus, we speculate that combining instructional techniques, such as lateral reading and MEL scaffolds, may facilitate students' deeper understanding of astronomy. We recognize that doing so may be counter to more traditional instruction (e.g., lecture), particularly at the undergraduate level where large survey courses dominate (e.g., Prather et al., 2009), but giving students more autonomy supportive instruction via cleverly designed scaffolding may result in active classroom environments that are more cognitively, behaviorally, and emotionally engaging (Lombardi et al., 2021). Effective science instruction often includes small group discussions involving negotiation and collaboration, essential scientific practices that may facilitate more critical evaluations (Ford, 2015; Governor et al., 2021). The use of these scaffolds may also help further students' understanding of the nature of evidence and explanations (Brickhouse et al., 2002). Although our research suggests that MEL scaffolds may be useful to promote students' more scientific evaluations and judgments, and help them to engage in productive scientific discussions, they are not intended to be a full curriculum. They are, in fact, relatively short lessons (~90 minutes) and intended to supplement and replace less effective activities. In short, we suggest that they are one of a suite of activities that may be used to help make astronomy instruction more effective.

7 Conclusion

Astronomy is a popular but challenging subject, due in part to the complexity and abstract nature of the topics, not to mention the presence of alternative conceptions and misinformation. Opportunities for students to evaluate explanations about astronomical topics can improve their learning. The results of the present study revealed a slight advantage of the baMEL, compared to the pcMEL, in promoting (a) deeper levels of evaluation between lines of evidence and alternative explanatory models; (b) plausibility shifts toward the scientific model; and (c) increased understanding of astronomy topics. This is consistent with previous studies (Klavon et al., 2021; Lombardi et al.,

2018a,b; Medrano et al., 2020) that reported an impact on knowledge gain after participating in MEL activities. Furthermore, we believe the evaluative practice incorporated in learning science can deepen students' scientific literacy, a major goal of many astronomy courses (Partridge and Greenstein, 2003).

8 Acknowledgements

We would like to thank Svetha Mohan, Josh Jaffe, and Jessica McLaughlin for their feedback, discussions, and support.

9 Funding

This work was supported by the U.S. 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.

References

Abdi, H. and Williams, L. J. (2010). Jackknife. In Salkind, N., editor, *Encyclopedia of research design*, volume 2, pages 655—-660. Sage, Thousand Oaks, CA, USA. https://doi.org/10.4135/9781412961288.n202.

Aretz, S., Borowski, A., and Schmeling, S. (2016). A fairy-tale creation or the beginning of everything: Students' pre-instructional conceptions about the Big Bang theory. *Perspectives in Science*, 10:46–58. https://doi.org/10.1016/j.pisc.2016.08.003.

Arthurs, L. and Templeton, A. (2009). Coupled collaborative in-class activities and individual follow-up homework promote interactive engagement and improve student learning outcomes in a college-level environmental geology course. *Journal of Geoscience Education*, 57(5):356–371. https://doi.org/10.5408/1.3544287.

Bailey, J. M., Coble, K., Cochran, G., Larrieu, D., Sanchez, R., and Cominsky, L. R. (2012). A multi-institutional investigation of students' preinstructional ideas about cosmology. *Astronomy Education Review*, 11(1). https://doi.org/10.3847/AER2012029.

Bailey, J. M., Girtain, C., and Lombardi, D. (2016). Understanding the formation of the Earth's Moon. *The Earth Scientist*, 32(2):11–16. https://www.nestanet.org/resources/Documents/Advocacy/TES/2015-2020/Summer16.pdf.

Bailey, J. M., Jamani, S., Klavon, T. G., Jaffe, J., and 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., and Dobaria, A. (2020). The origins build-a-MEL: Introducing a scaffold to explore the origins of the universe. *The Earth Scientist*, 36(3). https://www.nestanet.org/resources/Documents/Advocacy/TES/2015-2020/Fall20.pdf.

Bailey, J. M., Lombardi, D., Cordova, J. R., and Sinatra, G. M. (2017). Meeting students halfway: Increasing self-efficacy and promoting knowledge change in astronomy. *Physical Review Physics Education Research*, 13(2):020140. https://doi.org/10.1103/PhysRevPhysEducRes.13.020140.

Berg, C. (2014). Impacts of incorporating small-group active-learning activity modules on student achievement and attitudes in an introductory geology course. https://doi.org/10.13140/2.1.3897.7607.

Bondi, H. and Gold, T. (1948). The steady-state theory of the expanding universe. *Monthly Notices of the Royal Astronomical Society*, 108(3):252–270. https://doi.org/10.1093/mnras/108.3.252.

Braasch, J. L. G. and 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.

Brewe, E., Kramer, L., and O'Brien, G. (2009). Modeling instruction: Positive attitudinal shifts in introductory physics measured with CLASS. *Physical Review Special Topics - Physics Education Research*, 5(1). https://doi.org/10.1103/PhysRevSTPER.5.013102.

Brickhouse, N. W., Dagher, Z. R., Shipman, H. L., and Letts, W. J. (2002). Evidence and warrants for belief in a college astronomy course. *Science and Education*, 11(6):573–588. https://doi.org/10.1023/A:1019693819079.

Cardinot, A. and Fairfield, J. A. (2021). Alternative conceptions of astronomy: How Irish secondary students understand Gravity, Seasons, and the Big Bang. *Eurasia Journal of Mathematics, Science and Technology Education*, 17(4):em1950. https://doi.org/10.29333/ejmste/10780.

Ceyhan, G. D., Lombardi, D., and Saribas, D. (2021). Probing into pre-service science teachers' practices of scientific evaluation and decision-making on socio-scientific issues. *Journal of Science Teacher Education*, 32(8):865–889. https://doi.org/10.1080/1046560X.2021.1894762.

Chin, C. and Osborne, J. (2010). Students' questions and discursive interaction: Their impact on argumentation during collaborative group discussions in science. *Journal of Research in Science Teaching*, 47(7):883–908. https://doi.org/10.1002/tea.20385.

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

Coble, K., McLin, K. M., Bailey, J. M., Metevier, A. J., Peruta, C. C., and Cominsky, L. R. (2015). *The big ideas in cosmology*. Kendall Hunt Publishers / Great River Technology, Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Laurence Erlbaum Associates.

Dole, J. A. and Sinatra, G. M. (1998). Reconceptalizing change in the cognitive construction of knowledge. *Educational Psychologist*, 33(2-3):109–128. https://doi.org/10.1080/00461520.1998.9653294.

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.

Fraknoi, A. (2001). Enrollments in astronomy 101 courses. *Astronomy Education Review*, 1(1):121–123. https://doi.org/10.3847/aer2001011.

Friedman, A. (1922). Über die krümmung des raumes. *Zeitschrift für Physik*, 10(1):377–386. https://doi.org/10.1007/BF01332580.

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

Goodhue, D. L., Lewis, W., and 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.

Governor, D., Lombardi, D., and 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.

Hansson, L. and Redfors, A. (2006). Swedish upper secondary students' views of the origin and development of the universe. *Research in Science Education*, 36(4):355–379. https://doi.org/10.1007/s11165-005-9009-y.

Hayes, A. F. and Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But.... *Communication Methods and Measures*, 14(1):1–24. https://doi.org/10.1080/19312458.2020.1718629.

Henseler, J. and 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.

Hoyle, F. (1948). A new model for the expanding universe. *Monthly Notices of the Royal Astronomical Society*, 108(5):372–382. https://doi.org/10.1093/mnras/108.5.372.

Klavon, T. G., Mohan, S., Jaffe, J., Stogianos, T., Lombardi, D., and Governor, D. (2021). Scaffolding middle students' reasoning and learning about complex geoscience topics hydraulic fracturing and fossil evidence. *Journal of Geoscience Education*. in review.

Klosterman, M. L. and Sadler, T. D. (2009). 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. (2020). *WarpPLS User Manual: Version 7.0*. ScriptWarp Systems.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4):241–253. https://doi.org/10.3102/0013189x20912798.

Krstovic, M., Brown, L., Chacko, M., and Trinh, B. (2008). Grade 9 astronomy study: Interests of boys and girls studying astronomy at Fletcher's Meadow Secondary School. *Astronomy Education Review*, 7(2):18–24. https://doi.org/10.3847/AER2008017.

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

LaDue, N. D., McNeal, P. M., Ryker, K., John, K. S., and van der Hoeven Kraft, K. J. (2021). 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.

Larrain, A., Howe, C., and Freire, P. (2017). 'More is not necessarily better': Curriculum materials support the impact of classroom argumentative dialogue in science teaching on content knowledge. *Research in Science & Technological Education*, 36(3):282–301. https://doi.org/10.1080/02635143.2017.1408581.

Lelliott, A. and Rollnick, M. (2010). Big ideas: A review of astronomy education research 1974-2008. *International Journal of Science Education*, 32(13):1771–1799. https://doi.org/10.1080/09500690903214546.

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.

Lombardi, D., Bailey, J. M., Bickel, E. S., and Burrell, S. (2018a). 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., and Burrell, S. (2018b). 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., and Burg, C. (2016a). 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., and Young, N. (2016b). 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., and Sinatra, G. M. (2016c). 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., John, K. S., van der Hoeven Kraft, K. J., and and, J. L. D. (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., Sibley, B., and Carroll, K. (2013a). What's the alternative? using model-evidence link diagrams to weigh alternative models in argumentation. *The Science Teacher*, 080(05). https://doi.org/10.2505/4/tst13_080_05_50.

Lombardi, D. and Sinatra, G. M. (2010). College students' perceptions about the plausibility of human-induced climate change. *Research in Science Education*, 42(2):201–217. http://doi.org/10.1007/s11165-010-9196-z.

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

Lombardi, D., Sinatra, G. M., and Nussbaum, E. M. (2013b). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction*, 27:50–62. https://doi.org/10.1016/j.learninstruc.2013.03.00.

Mason, L., Ariasi, N., and Boldrin, A. (2011). Epistemic beliefs in action: Spontaneous reflections about knowledge and knowing during online information searching and their influence on learning. *Learning and Instruction*, 21(1):137–151. https://doi.org/10.1016/j.learninstruc.2010.01.001.

McGrew, S. (2020). Learning to evaluate: An intervention in civic online reasoning. *Computers & Education*, 145:103711. https://doi.org/10.1016/j.compedu.2019.103711.

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

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

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

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

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

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

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

Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., Pierczynski, M., and Allen, M. (2017). Teach-

ers' instructional adaptations: A research synthesis. *Review of Educational Research*, 88(2):205–242. https://doi.org/10.3847/aer2003016.

Partridge, B. and Greenstein, G. (2003). Goals for "astro 101:" report on workshops for department leaders. *Astronomy Education Review*, 2(2):46–89. https://doi.org/10.3847/aer2003016.

Patall, E. A., Pituch, K. A., Steingut, R. R., Vasquez, A. C., Yates, N., and 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.

Plummer, J. D. and Maynard, L. (2014). Building a learning progression for celestial motion: An exploration of students' reasoning about the seasons. *Journal of Research in Science Teaching*, 51(7):902–929. https://doi.org/10.1002/tea.21151.

Prather, E. E., Rudolph, A. L., Brissenden, G., and Schlingman, W. M. (2009). A national study assessing the teaching and learning of introductory astronomy. Part I. the effect of interactive instruction. *American Journal of Physics*, 77(4):320–330. https://doi.org/10.1119/1.3065023.

Prather, E. E., Slater, T. F., and Offerdahl, E. G. (2002). Hints of a fundamental misconception in cosmology. *Astronomy Education Review*, 1(2):28–34. https://doi.org/10.3847/AER2002003.

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.

Reeve, J. and 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.

Roemmele, C. (2017). *Unearthing geologic blindness: Undergraduate students attitude and conceptual understanding of Geology*. PhD thesis, Purdue University.

Sinatra, G. M., Heddy, B. C., and 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. and 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* > .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.

Tabachnick, B. G. and Fidell, L. S. (2007). *Using multivariate statistics (5th ed.)*. Pearson Education., 5 edition.

Trouille, L. E., Coble, K., Cochran, G. L., Bailey, J. M., Camarillo, C. T., Nickerson, M. D., and Cominsky, L. R. (2013). Investigating student ideas about cosmology III: Big Bang theory, expansion, age, and history of the universe. *Astronomy Education Review*, 12(1). https://doi.org/10.3847/AER2013016.

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.

Wasserstein, R. L., Schirm, A. L., and 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.