Using the Critical Questions Model of Argumentation for Science Teacher Professional Learning and Student Outcomes

E. Michael Nussbaum, Lixian Tian, Michael S. Van Winkle, Harsha N. Perera, LeAnn G. Putney, and Ian J. Dove
nussbaum@unlv.nevada.edu, tianlixian2011@gmail.com, vanwinkl@unlv.nevada.edu, harsha.perera@unlv.edu, leann.putney@unlv.edu; ian.dove@unlv.edu
University of Nevada, Las Vegas

Kris Carroll, Southern Nevada Regional Profession Development Program, carrollk@rpdp.net

Abstract: The Critical Questions Model of Argument Assessment (CQMAA) poses questions for evaluating arguments. Seven secondary science teachers completed a year-long professional learning program on using the CQMAA with argument pedagogy. Participants designed and implemented three argument-based lessons. We surveyed (pre/post) teacher and student confidence with using argument pedagogy; students also completed a test assessing their use of critical questions (CQs). We found improvements in teacher and student confidence, and students also became better at critiquing premises (especially analogies). Results regarding evidence critiques and considering alternative explanations were mixed. Also, the teachers reported that the critical questions (CQs) deepened students' reasoning, but some CQs were more useful than others. The research suggests that developing teachers' pedagogical content knowledge regarding scientific argumentation is a lengthy process, and that more (video) examples of expert teachers using argument pedagogy are needed, but the CQs are useful tools for both teachers and students.

Keywords: Argumentation, Science Education, Professional Development, Critical Thinking

A core Science and Engineering Practice described in *A Framework for K-12 Science Education* (NRC, 2012) and written throughout the Next Generation Science Standards (NGSS Lead States, 2013) is "Engaging in Argument from Evidence." For middle school, the practice element expectations include (a) constructing arguments supported by empirical evidence and scientific reasoning to support or refute an explanation or a model for a phenomenon or a solution to a problem, and (b) respectfully providing and receiving critiques "about one's explanations, procedures, models, and questions…" (NGSS Lead States, 2013, Appendix F). The first of these has received more attention by professional educators than the second (NSTA, 2019), so to address this imbalance, we developed a model of argument critique known as the Critical Questions Model of Argument Assessment (CQMAA). In a year-long research project, we provided professional learning to a group of mostly middle-school science teachers. The teachers designed lessons that supported both practice elements, including mapping claim-evidence-reasoning (CER) components of arguments onto critical questions and using the critical questions to provide and receive critiques. We investigated which critical questions the teachers used, how they used graphic organizers to incorporate CQs into their lessons, and the effects on teacher confidence and student outcomes.

Theoretical framework

The first performance expectation reflects a model of scientific argumentation first proposed by McNeil and Krajcik (2007), albeit rooted in a more complex model developed by Toulmin (1958). McNeil and Krajcik's CER model specifies three argument components: the claim (or conclusion), the evidence, and reasoning that connects the evidence to the claim. It is common for teachers to provide students with worksheets and graphic organizers to help them generate these components (Belland et al., 2008). One limitation of this approach, however, and of Toulmin-based argument approaches in general, is the lack of standards addressing the quality of the components, and more generally for critiquing arguments (Nussbaum, 2011). Some researchers have addressed these limitations by adding a "refutations" component to the CER model (Erduran et al., 2004), using evaluative prompts for oral discussions (Sampson & Grooms, 2010), and having science classroom communities develop "epistemic standards" for the reliability of evidence (Pluta et al., 2011). Chinn and colleagues also developed graphic organizers known as model-evidence link diagrams (MELs) that ask students to assess two competing scientific models by specifying whether certain pieces of evidence support or contradict each model (Chinn & Buckland, 2012).

We built upon these efforts by developing an additional set of tools specifically designed for scaffolding students' abilities to critique arguments (using the CQMAA), and to provide support for the reasoning component of the CER model (not just the evidence component). The tools are also intended to be used to support a form of oral discourse known as *collaborative argumentation* (Andriessen, Baker, & Suthers, 2003), in which student work together to construct and critique arguments. Collaborative argumentation has been found to promote deeper engagement and understanding of scientific concepts (Alexopoulou & Driver, 1996; Asterhan & Schwarz, 2007).

The Critical Questions Model of Argument Assessment (CQMAA)

The CQMAA (Dove & Nussbaum, 2018) is rooted in the work on argument schemes by various philosophers (e.g., Verheij, 2003), the most prolific of whom has been Walton (1996). An argument scheme is a type of argument, and attached to each scheme is a set of *critical questions* that should be used to assess the strength and cogency of an argument reflecting that scheme. Walton et al. (2007) have identified over 60 such schemes; however, we believe—based on our personal experience—that teaching students to classify arguments into schemes is often unwieldy, not only because of the cognitive load involved, but because many arguments do not fit exactly into the scheme patterns identified by Walton and others. Noting, however, that many of the critical questions associated with different schemes are similar to one another, we developed a list of eight general critical questions (CQs) that can be used to evaluate most arguments (see Nussbaum & Dove, 2018). These questions are shown in Table 1.

Except for the first, individuals can ask these questions in any order, or use just a subset of the questions. To spur deeper collaborative argumentation, participants explore one or more of the CQs. If a CQ can be answered affirmatively, it strengthens the argument, whereas a negative answer weakens it.

It should be noted that all of the CQs can be applied to scientific arguments except the set in #8, which only apply to arguments that one should do something. Such arguments involve value judgments which are not empirical. Nevertheless, this set of CQs is applicable to socioscientific issues, where competing values must be weighed, or design/engineering problems which involve evaluating "competing design solutions based on jointly developed and agreed-upon design criteria" (NSTA, 2019).

Finally, in using the CQMAA, we encourage teachers and researchers to redesign the wording of the CQs to make them accessible to students at different grade levels. Table 1 reflects our attempt to make the initial questions (from Nussbaum & Dove, 2018) more accessible to middle-school students. A teacher can also propose additional, more specific CQs, for example about the reliability of a source when discussing the more general CQ regarding quality of evidence (#2 in Table 1).

Table 1: CQMAA critical questions

Number	Critical Questions
1.Structure	Is there an argument here? Can I underline the claim? Can I add a star to the evidence? Can I bracket the reasoning?
2.Evidence	Is there evidence? How good is the evidence?
3.Reasons	Are any of the reasons untrue or incorrect?
4.Accurate	Does the argument use or connect with accepted scientific laws?
5.Coherence/ Reasoning	Do the parts of the argument make a path you can follow or are there missing steps?
6.Alternatives	Are there other claims or conclusions that also fit the evidence? Can you rule out other or competing claims or conclusions?
7.Completeness	What is missing or weak in the argument?
8.Trade-offs	For engineering design or socioscientific arguments: Are there trade-offs (getting something at the expense of another)? Are any of the reasons on one side more important than those on the other? Can you design a creative solution?

Graphic organizers

We developed two graphic organizers that educators can use to incorporate the teaching of critical questions into their lessons. These include:

- *Argument Analysis Mapping Tool*: Students generate CER components in the left column and then answer CQs from Table 1 in the right column. We also developed a guide to using the mapping tool.

- *Evidence-to-Hypothesis Argument Scheme (EHAS) Diagrams*. Requires students to fill in three boxes on hypothesis, reasoning, and evidence (explaining how a hypothesis would result in the evidence). A peer then answers two CQs that consolidate some shown in Table 1. The two CQs address (a) understanding and improving the explanation, and (b) whether the evidence and reasoning are correct.

We also used: MEL diagrams (described previously) to help students think about alternative explanations (CQ #6 in Table 1), an Explanation Tool to help students construct explanations from reasoning and evidence, Argument Vee Diagrams (Nussbaum, Dove, et al., 2019) with the CQs in #8 in Table 1, and accountable talk cards (Michaels & O'Connor, 2012).

Theory of change

Many science teachers have had little if any training in either argument theory or argument pedagogy, and also may have gaps or misconceptions in their content knowledge (Sampson & Blanchard, 2012). It is therefore important to develop both their content knowledge (CK) and pedagogical content knowledge (PCK) involving argumentation regarding subject matter. There were several important aspects of our professional learning (PL) design that addressed this. First, the PL was intensive, long-term, and tied to classroom practice, which are critical features of effective teacher learning (Yoon et al., 2007). The PL consisted of a phenomenon-driven 10-day summer institute, eight follow-up sessions during the school year (monthly group meeting of 2.5 hours each), and observation and feedback sessions by a discourse coach three times during the school year (one per quarters 2-4). Discourse coaches were either teachers with more experience in dialogic pedagogy, instructional coaches, or in two cases, employees of the district's professional learning office.

Second, teachers experienced dialogic and argument pedagogy first-hand, as students trying to make sense of phenomenon. Doing so has been found to make teachers more appreciative of the affordances and constraints underlying a particular tool or technique, and to help build both CK and PCK (Borko, 2004). The summer institute was organized around having teachers use the various tools described above as students, and collecting evidence to make sense of phenomena related to forces and motion, core ideas most were expected to teach that fall.

Third, during the school year, the teachers engaged in designing and implementing three (week-long) learning sequences using argumentation, one per quarter. Teachers and their coaches used a lesson development and review form to identify where in a lesson students' would generate and evaluate arguments, engage in discourse, and use CQs. After observing a portion of each learning sequence, the coaches used an observation form and a semi-structured feedback form that was utilized to set short-term goals for further teacher development. The follow-up sessions were used in part for collaborative lesson planning and for discussing with the other participants common implementation challenges. These collaborative engagement points and inquiry practices have been found to be vital ingredients of effective teacher learning (Wilkinson et al., 2017).

Research goals

The first goal of this research was to determine whether teachers could incorporate the CQs into their lessons and would find them useful in generating richer student oral and written discourse and reasoning. The second goal was to evaluate our PL program: What tools would teachers use from the project, would their confidence in argument pedagogy increase, what challenges would they face, and how could the PL be improved? The third goal was to examine the effect of the PL program on student outcomes.

Methodology

Participants

There were 17 participants in the summer institute but, because of attrition for various reasons, the final sample consisted of six middle-school science teachers and one high school teacher, plus seven coaches. Of the teachers, 57% were eighth-grade teachers, 79% female, and 79% white. They worked at schools of various SES levels. The final student sample consisted of 230 students from whom parental consent and student assent were obtained.

Instruments and study design

The study used a mixed-methods QUAN-qual design. Quantitative measures consisted of:

- A 30-item teacher survey (using 5-pt. Likert scales) assessing (a) confidence in using argument pedagogy (e.g., planning lessons, prompting students to elaborate), (b) confidence in engaging in argument (e.g., arguing for a position), and (c) beliefs about argument pedagogy (general and self-referenced), such as "arguing leads to more understanding." These subscales were validated in a prior study (Nussbaum, Van

Winkle, et al., 2019), with internal consistencies of .91, .91, .85, and .87, respectively (McDonald's Ω). The survey was administered three times, at the start of the summer institute, at the beginning of the third school quarter, and at the end of the project. There were also two open-ended items where teachers could elaborate on their answers.

- A 20-item teacher survey using 5-pt. Likert scales (from Nussbaum, Van Winkle, et al., 2019) on the extent that teachers used various dialogic practices. It was administered with the teacher confidence survey. There were three subscales: "Teaching Critique of Arguments/Concepts" (e.g., planning lessons involving students critiquing one another's arguments), "Teacher Talk Moves" (e.g., rephrasing student comments), and "Assessment/Writing" (e.g., having students complete graphic organizers related to argumentation). Ω's were .83, .92, .89, respectively.
- Student scores on the Argument Evaluation Test (AET), developed for this study. Students completed the test in the same class session as the confidence survey. We developed two different test forms, counterbalancing the administration. Each form consisted of two scenarios, each containing two opposing arguments on a topic (e.g., global climate change, cause of the seasons). Students had to choose which of the two arguments was weakest (and to explain why), to describe as many flaws in the weaker argument as possible, and to explain which of these flaws is the most important. The intent was to assess whether students were applying the criteria underlying the CQs in identifying flaws. The scenarios were geared toward seventh-grade topics so that most of the students (i.e., those in eighth or eleventh grade) presumably had some knowledge of the topics. There were also three multiple-choice questions per scenario assessing whether students could identify the CER components (corresponding to CQ #1). Each test was scored by three raters using a rubric with categories corresponding to the CQs in Table 1 plus other categories that emerged during scoring, with differences resolved through discussion. Students received one point by category for each flaw identified, plus an additional point if their criticism was elaborated. The average intraclass correlation among raters was used to measure of reliability, with reliabilities ranging from .70 to .89, except for one category (identifying logical gaps). This category (ICC = .38) was removed from the analysis.
- A 30-item student confidence survey (using 5-pt. Likert scales) measuring student self-efficacy for collaborative argumentation (e.g., "I feel comfortable offering my ideas during a discussion"). Where possible, each item was paired with another measuring the same construct, with reliabilities (i.e., correlations between pairs) ranging from .68 (for evaluating arguments) to .78 (for generating rebuttals). Items pertaining to writing, identifying weaknesses, alternative explanation, and "relating arguments to the big question" were unpaired.
- We used as a covariate students' prior-year scale scores in reading on a state-mandated test.

Qualitative data consisted of the following:

- Transcripts of semi-structured exit interviews that were conducted of each participant and their coach. Participants were asked about their confidence levels, tool use, perceptions of student improvement, and how the PL program could be modified. Thematic analysis was applied to these data.
- The coaching feedback forms for each lesson.
- Comments from a teacher focus group (conducted mid-year) on what was working and what could be improved regarding the PL design.

Results

Extent of tool use

Based on the nonparametric Page test (b = 6), extent of use improved over time on all three subscales (p's < .05). The average scores during the PL program (i.e., averages of the second and third administrations) were Critiquing Arguments & Concepts ($M = 3.48$, $SD = 0.55$), Use of Talk Moves ($M = 3.81$, $SD = 0.42$), and Writing/Assessment ($M = 3.64$, $SD = 0.41$). A rating of 3 reflected using "sometimes" and 4 "often." The open-ended responses indicated that mentees became more cognizant of various functions of argumentation (e.g., in promoting engagement or understanding).

The coaches completed statistics for each lesson observed, specifically on the tools utilized and how well each was used. Based on these data, we determined that most of the learning sequences (92%) used a graphic organizer and 50% incorporated critical questions (most commonly through use of the mapping tool). Only two of the teachers consistently used the same graphic organizer in all learning sequences (mapping tool, or EHAS Diagram); most of the other teachers indicated that they wanted to try out different tools.

The exit interview data indicated that 57% of the teachers used MEL diagrams, half consistently and half in only one learning sequence. One teacher consistently used the EHAS Diagram, and one used the vee diagram in one lesson. Two teachers used the explanation tool, and three used the argumentation mapping tool (which used CQs to evaluate CER-structured arguments). Teachers seemed to appreciate the concrete and structured nature of the graphic organizers; as one stated "I think that's what I was mostly looking for…something concrete that I could actually hold onto and use…something I can add to."

Use and value of critical questions

Regarding the critical question, teachers varied in the number of CQs emphasized in their lessons and in the graphic organizers used. Some addressed all of those in Table 1, some just one or two (typically related to evidence). One teacher who used the argument mapping tool noted that students only had time and attention to focus on the first four CQs in Table 1 and, as will be noted later, the effects on student outcomes associated with these questions were weaker.

The data from the exit interviews indicated that all but one of the teachers found the CQs valuable. Two teachers identified numerous advantages of CQs. These teachers reported that the CQs provided a structure for analysis, helped keep small-group discussions more focused and on task, made the students think, and improved the quality of CER arguments. One also noted that the CQs provided students with terms, such as counterevidence or counterarguments, to better express their ideas and critiques. Her discourse coach noted, "So [the students are] thinking, but it felt like a lot of them just didn't have that way to share their thinking and the critical questions or just practicing using those critical questions really brought it out."

Two other teachers viewed the CQs positively but with caveats, specifically that the number should be limited (this teacher preferred the two on the EHAS Diagram) or that students had to be developmentally prepared for specific CQs. Two other teachers only incorporated the CQs into one of their learning sequences, and indicated that the teacher had to build up their confidence to use them. One teacher did not use CQs, indicating that they were too complex for students.

Teacher confidence and quality of tool use

There were significant increases over time in the teachers' confidence in implementing argument pedagogy (Page's L = 81.5, p < .01), but not for personally engaging in arguments or in beliefs about collaborative argumentation. One teacher, Karen, articulated in her exit interview what other teachers similarly noted regarding their confidence levels, "I feel like I have a better foundation now than I had before…this was really the trial year, I think, with discourse in the classroom. So next year I would like to make it more my expertise."

The survey results regarding improvement in pedagogy were further confirmed by data drawn from the coaching reflection forms on quality of tools use. Table 2 shows that the coaches' mean ratings improved over time, from about a 3 in Qtr. 2 (reflecting "purpose somewhat evident") to 6's in Qrt. 4 ("use very effectively with good sequence fit and evident purpose)."

Table 2: Coaches' mean quality ratings of use of graphic organizers and critical questions

Tool category	Qtr. 2	Qtr. 3	Qtr 4	Overall *M* (*SD*)
Graphic organizers	2.8	5.0	6.0	4.45 (1.61)
Critical questions	3.0	5.3	6.0	4.43 (1.40)

Student outcomes

Survey outcomes

The student data were analyzed using structural equation growth modeling with full information maximum likelihood (FIML) estimation. This approach allowed us to use the data for students who completed the pretest but not the posttest or vice versa, as that information was still useful in assessing and controlling for form effects. Mean-centered reading scores were used as a covariate to better meet the missing-at-random assumptions of FIML estimation. Robust standard errors clustered by teacher were used to control for the nested nature of the data (students nested in teachers). Latent slopes were estimated for each student, representing the improvement of each student from pre to post while controlling for the test form used (in the case of the AET) and the other predictors. In the following, $\hat{\beta}$ represents the average of the individual student slope estimates.

For the student survey data (N = 209), there were significant increases in students' confidence in providing evidence ($\hat{\beta}$ = 0.20, p = .002), knowledge of what makes a good argument ($\hat{\beta}$ = 0.19, p < .001), evaluating arguments ($\hat{\beta}$ = 0.21, p = .003), generating rebuttals ($\hat{\beta}$ = 0.22, p < .001), writing argument critiques ($\hat{\beta}$= 0.22, p < .001), identifying weakness in an argument ($\hat{\beta}$ = 0.17, p = .014) and relating arguments to the big question ($\hat{\beta}$ = 0.22, p = .002). Significant results (p < .05) were also found for generating alternative explanations ($\hat{\beta}$ = 0.27, p =.018) and asking questions during oral discussions ($\hat{\beta}$ = 0.15, p = .027). Overall, student confidence in critiquing arguments increased.

Only two of the categories assessed by the AET (N = 184) showed significant increases: critiquing an analogy ($\hat{\beta}$ = 0.45, p < .001) and challenging the truth of a premise ($\hat{\beta}$ = 0.12, p = .001). Results were nonsignificant for identifying alternative explanation or counterevidence ($\hat{\beta}$ = -0.12, p = .11), and CER component identification ($\hat{\beta}$= -0.40, p = .10), and negative for two categories: evidence critique ($\hat{\beta}$ = -0.72, p = .002) and identifying a premise as unsupported ($\hat{\beta}$ = -0.11, p = .04). The latter was marginally significant, and could reflect a Type I error given the number of statistical tests performed.

Possible explanation of student survey results

In summary, students became more confident in critiquing arguments, and showed greater competency in critiquing analogies and, to a lesser degree, challenging premises. Both of these involve stimuli that are more visually and cognitively salient, in that an analogy or premise is given to a student, whereas the lack of support or the possibility of an alternative explanation or counterevidence is not when attention is focused on the weaker argument. These competencies may be more cognitively complex. Although student confidence for providing alternative explanations increased, possibly due to the use of MELs or from the CQs, students may have needed greater practice in more diverse settings for this skill to transfer to a test situation involving lengthy text.

The evidence critique effect was highly significant and negative, and at odds with the confidence finding. One explanation, suggested by the exit interviews, is that different teachers focused on different facets of evidence (e.g., supporting statements with evidence, having enough evidence, having strong evidence, describing the evidence in more detail, determining whether evidence supported or contradicted a statement), and the AET measures were possibly not broad and sensitive enough to detect these differences. This might explain a null effect but not a negative effect. It is possible that students with high pretest scores chose other types of faults on which to focus during the posttest, which would then make an otherwise null effect negative.

The null effect of identifying the CER components was a little surprising, given the widespread use of graphic organizers using a CER structure, but one teacher noted that students often wrote evidence in the reasoning box and vice versa. Another two teachers were unsure of the difference between reasoning and evidence and, while we addressed this during the PL, more systematic instruction may be needed.

Challenges and recommendations for improving PL design

Although the participants found experiencing use of the argumentation tools (including CQs) valuable in the summer institute, planning learning sequences integrating these tools proved time consuming and challenging. Teachers had to find appropriate topics and questions consistent with the NGSS standards and district curriculum, sequence lesson activities appropriately while attending to conceptual development, scaffold argumentation skills, and maintain discussion norms conducive to critique. In conceptual terms, teachers needed to acquire and integrate pedagogical content knowledge (PCK) about content with PCK about argumentation. During the first year they tried out various tools (and CQs tied to those tools), but one participant described these as "baby steps" and another claimed that an additional year of intense professional learning was needed.

Regarding other possible design changes, during a focus group held mid-year, the teachers requested examples of expert teachers conducting whole-class and small-group argumentation lessons, and while we showed two videos of expert teachers (based on Hunt & Minstrell, 1994, and Osborne, Donovan, Henderson, MacPherson, & Wild, 2016), we concluded more videos could be shown earlier in the PL program. Mentee observation of coaches would also be worthwhile. Except for the summer institute, most teachers had not experienced dialogic pedagogy as students, and several teachers claimed during the exit interviews that they needed to see what this type of pedagogy looked like. Giving participants more guidance on productive topics and questions in the curriculum, and of possible activities (like those contained in Osborne et al., 2016), were also suggested. Also one participants reported that it was more productive to have students complete the second column of the mapping tool before the first (i.e., to use CQs to construct the components of a CER diagram).

Finally, the majority of teachers indicated during the exit interviews that they often did not have sufficient time to implement the learning sequences without falling behind in curriculum coverage, so this is an issue that needs to be addressed in future PL projects

Discussion

We introduce in this paper a new set of tools, Critical Questions (CQs), to strengthen science students' ability to critique arguments and to make student written and oral arguments richer and more complex. We found supporting evidence for the utility of this approach, especially from the interview data and student survey data (student confidence in critiquing arguments increased over time). Some support was also provided by student performance on the argument evaluation test (AET), especially in regards to critiquing premises and analogies. Using CQs to improve reasoning on, or about, CER-structured arguments, appeared especially promising.

There were mixed results regarding whether students became better at critiquing evidence, with a positive effect on the confidence survey and a negative effect on the AET. A more systematic approach to evidence critique is likely needed, such as the one used by Pluta et al. (2011).

Our results have some implications for learning progressions in argumentation (e.g., Osborne et al., 2016). CQs relating to more visually and cognitively salient issues (such as the truth of premises) may be initially more accessible to students than CQs related to considering alternative models. Teachers did use MEL diagrams to make alternative models more salient, but there were no effects on the AET. The AET items provided alternative models but were text heavy and so perhaps did not make these models sufficiently salient and simple enough for most students to process. An alternative account is that most students did not attend much during the instruction to the CQ related to alternative explanations as it was presented further down the list than the one pertaining to premise critique. Further research is needed to test these competing accounts.

Teacher confidence in using argument pedagogy also increased, although there were differences in the number of CQs that teachers emphasized in their lessons. There are a number of factors to which these differences could be related, such as variation in (a) teachers' understanding of the CQs, (b) their PCK of scientific argumentation, (c) their beliefs about their students' developmental readiness, and (d) their choice of what graphic organizers to use in their learning sequences.

Teacher lesson planning was found to be more complex and time consuming than we expected. More generally, we found acquiring adequate PCK of scientific argumentation to be a lengthy process. However, our research shows that critical questions can be incorporated into relatively simple scaffolds to enhance teachers' and students' understanding of complex scientific argumentation.

References

Alexopoulo, E., & Driver, R. (1996). Small-group discussion in physics: Peer interaction modes in pairs and fours. *Journal of Research in Science Teaching, 33*(10), 1099-1114. https://doi.org/10.1002/(SICI)1098-2736(199612)33:10<1099::AID-TEA4>3.0.CO;2-N

Andriessen, J., Baker, M., & Suthers, D. (2003). Argumentation, computer support, and the educational context of confronting cognitions. In Andriessen, J., Baker, M., & Suthers, D. (Eds.), *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments* (pp. 1-25). doi.org/10.1007/978-94-017-0781-7_1

Asterhan, C. S. C., & Schwarz, B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of Educational Psychology, 99*, 626-639. doi.org/bkn25r

Belland, B. R., Glazewski, K. D., & Richardson, J. C. (2008). A scaffolding framework to support the construction of evidence-based arguments among middle school students. *Educational Technology Research and Development*, *56*, 401-422. doi:10.1007/s11423-007-9074-1.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain*. Educational Researcher, 33*(8), 3-15. doi.org/10.3102/0013189X033008003

Chinn, C. A. & Buckland, L. A. (2012). Model-based instruction: Fostering change in evolutionary conceptions and in epistemic practices. In K. S. Rosengren, E. M., Evans, S. Brem, & G. M. Sinatra (Eds.). *Evolution challenges: Integrating research and practice in teaching and learning about evolution* (pp. 211-232. Oxford University Press. doi:10.1093/acprof:oso/9780199730421.001.0001

Dove, I., & Nussbaum, E. M. (2018). The critical questions model of argument assessment. In S. Oswald & D. Maillat (Eds.), *Argumentation and Inference: Proceedings of the 2nd European Conference on Argumentation*, Fribourg 2017 (Vol. 2, pp. 263-280). London, England: College Publications.

Erduran, S., Simon, S. and Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Science Education, 88*, 915–933. doi.org/10.1002/sce.20012

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 51-74). Cambridge University Press.

McNeil, K. L., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 233-265). Taylor and Francis. doi.org/10.4324/9780203810057

Michaels, S., & O'Connor, C. (2012). *Talk science primer*. Cambridge, MA: TERC. http://inquiryproject.terc.edu/shared/pd/TalkScience_Primer.pdf

National Research Council (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. The National Academies Press. doi.org/10.17226/13165.

National Science Teachers Association (2019). *Science and Engineering Practices.* Retrieved from https://ngss.nsta.org/Practices.aspx?id=7

NGSS Lead States (2013). *Next Generation Science Standards: For states, by states*, Appendix F: Science and engineering practices. National Academies. http://www.nextgenscience.org/

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

Nussbaum, E. M., & Dove, I. J. (2018, August). *The Middle Way: The Critical Questions Model of Argument Assessment*. Poster session presented at the biennial meeting of SIG 26 (Argumentation, Dialogue, and Reasoning) of the European Association for Research on Learning and Instruction, Hebrew University, Jerusalem, Israel.

Nussbaum, E. M., Dove, I. J., Slife, N., Kardash, C. M., Turgut, R., & Vallett, D. (2019). Using critical questions to evaluate written and oral arguments in an undergraduate general education seminar: A quasi-experimental study. *Reading and Writing, 3*2, 1531–1552. doi.org/10.1007/s11145-018-9848-3

Nussbaum, E. M., Van Winkle, M. S., & Perera, H. N. (2019, April). *Surveys for evaluating teachers' use, beliefs, and self-efficacy with classroom argumentation in secondary (science) classrooms*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Osborne, J. F., Donovan, B. M., Henderson, J. B., MacPherson, A. C., & Wild, A. J. (2016). *Arguing from evidence in middle school science: 24 activities for productive talk and deeper learning.* Corwin.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S.-Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*(6), 821-846. doi.org/10.3389/fpsyg.2019.01414.

Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching, 48*(5), 486-511. doi/org/10.1002/tea.20415

Sampson, V., & Blanchard, M. R. (2012). Science teachers and scientific argumentation: Trends in views and practice. *Journal of Research in Science Teaching, 49*(9), 1122-1148. doi:10.1002/tea.21037

Sampson, V., & Grooms, J. (2010). GENERATE AN ARGUMENT: An instructional model. *The Science Teacher, 77*(5), 32-37. https://eric.ed.gov/?id=EJ888144

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning.* Erlbaum.

Walton, D., Reed, C., & Macagno, F. (2007). *Argumentation schemes*. Cambridge University Press.

Wilkinson, I. A. G., Reznitskaya, A., Bourdage, K., Oyler, J., Glina, M., Drewry, R., Kim, M.-Y., & Nelson, K. (2017). Toward a more dialogic pedagogy: Changing teachers' beliefs and practices through professional development in language arts classroom. *Language and Education, 31*(1), 65-82.

Verheij, B. (2003). Dialectical argumentation with argumentation schemes. *Artificial Intelligence and Law, 11*(2-3), 167-195. doi>*10.1023/B:ARTI.0000046008.49443.36*

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. U.S. Department of Education. ies.ed.gov/ncee/edlabsies.ed.gov/ncee/edlabs/regions/southwest/pdf/rel_2007033.pdf