# Measuring pedagogical content knowledge of argumentation through the development of a teacher argumentation assessment

Katherine L. McNeill[1], Maria Gonzalez-Howard[1], Rebecca Katsh-Singer[1] and Suzanna Loper[2]

Boston College[1]

Lawrence Hall of Science, University of California, Berkeley[2]

contact info:

Katherine L. McNeill

Lynch School of Education, Boston College

140 Commonwealth Avenue, Chestnut Hill, MA 02467

Phone: 617-552-4229

Fax: 617-552-1840

kmcneill@bc.edu

Reference as:

McNeill, K. L., Gonzalez-Howard, M., Katsh-Singer, R. & Loper, S. (2014, March). *Measuring pedagogical content knowledge of argumentation through the development of a teacher argumentation assessment.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Pittsburgh, PA.

**Abstract**

Argumentation is a key practice of science that has also been included in recent reform documents and science standards as essential for k-12 instruction. Despite the recent emphasis on argumentation, little work has focused on teachers' knowledge of argumentation. Pedagogical content knowledge (PCK) encompasses a variety of complex knowledge essential for effective teaching including knowledge of students' conceptions and knowledge of instructional strategies. The development of a high quality assessment for teachers' PCK of scientific argumentation is important to better assess the needs of teachers as well as to evaluate the quality of their teacher education experiences. We present our initial efforts to conceptualize, develop and pilot a measure of teachers' PCK of argumentation. Our development and piloting process builds off the model proposed by Hill and her colleagues (2004; 2008), which includes conceptualization of the domain, design of items, pilot testing items and cognitive interviews. In this paper, we present the design of vignettes that use samples of student writing and classroom transcript to assess teachers' PCK of argumentation using both multiple-choice and open-ended items. In addition, we share the results from our pilot test with 103 teachers, cognitive interviews with 24 teachers, and feedback from 10 advisors. Although this is ongoing work, we feel that our lessons learned from these initial efforts offer important implications for the field in terms of others looking to support or assess teachers' PCK not only for argumentation, but also other scientific practices. Our work suggests the importance of using scientific practice related item choices as distractors, the challenge of designing answer choices that assess a deep understanding of the scientific practice rather than surface level features, the potential of using vignettes in PCK assessments, and the greater challenges of assessing the dialogic aspects of argumentation compared to structural components.

## Measuring pedagogical content knowledge of argumentation through the development of a teacher argumentation assessment

Scientists engage in argumentation in which they construct new knowledge of the natural world through the critique and revision of ideas within the scientific community (Osborne, 2010). Recent reform documents (National Research Council [NRC], 2012) and national standards (Achieve, Inc., 2013) call for students to engage in this same practice in k-12 classrooms. With the current focus on scientific practices in the Next Generation Science Standards (NGSS) (Achieve Inc., 2012), teachers need greater support for the scientific practices, such as argumentation, due to the novelty of having such practices explicitly incorporated within science standards (NRC, 2012). As a field we are beginning to design supports for teachers around the scientific practices, including professional development (Moon, Passmore, Reiser & Michaels, in press; Wilson, 2013) and educative curriculum materials (Loper, McNeill, Peck, Price & Barber, 2014). To assess the strengths and weaknesses of these teacher education experiences as well as to provide valuable information for teacher educators designing such experiences, we need measures that target teachers' pedagogical content knowledge (PCK) of scientific practices. PCK is subject matter knowledge for teaching that includes both an understanding of students' conceptions and appropriate instructional strategies that are important for effective classroom instruction (Shulman, 1986). Understanding teachers' PCK is essential for developing teacher education programs that meet teachers' needs throughout their careers (Schneider & Plasman, 2011).

Although this work is ongoing, we chose to write about the assessment development at this juncture, because we feel that we have important lessons learned to offer the field in terms of assessing and supporting teachers' PCK of scientific practices, such as argumentation. This manuscript follows a unique structure. We begin by providing our theoretical framework, which informed our conceptualization of scientific argumentation and our model for assessing and supporting PCK of argumentation. We then describe the seven steps in our PCK assessment development process to date, focusing on one vignette and corresponding assessment items to illustrate the process. Next we present four overarching lessons learned from this work using other assessment items and data from the pilot testing, cognitive interviews and advisors' feedback to illustrate these lessons. Finally, we discuss implications and potential directions for future work.

## Theoretical Framework

### *Argumentation*

Argumentation is a complex scientific practice that consists of both a structural focus including the types of justifications that our valued within the scientific community and a dialogic focus on the social process in which scientists interact (Jiménez-Aleixandre & Erduran, 2008). Although scientists support claims using a variety of kinds of justifications, specific justifications, such as empirical evidence, are valued over others (Sandoval & Cam, 2011). Consequently, a particular structure is frequently utilized in science. Similar to other science education researchers (Osborne, Erduran & Simon, 2004; Sampson & Clark, 2008), we adapted Toulmin's (1958) model of argumentation, which also aligns with the language in NGSS. Specifically, at the middle school level, NGSS (Achieve, Inc., 2013) states that students should, "Construct, use, and/or present an oral and written argument 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." (Appendix F, p. 29 – italics added). As such, we see the structure of a scientific argument consisting of a claim supported by evidence and reasoning (McNeill, Lizotte, Krajcik & Marx, 2006). Evidence includes scientific data, such as observations and measurements, while reasoning explains why the evidence supports the claim using disciplinary core ideas. When constructing arguments, teachers can have difficulty supporting their claims with appropriate evidence and reasoning (Sampson & Blanchard, 2012). Furthermore, when analyzing samples of student work, both in writing and in classroom discussions, teachers can have difficulty assessing students' reasoning as well as determining appropriate instructional supports to help improve students' reasoning (McNeill & Knight, 2013). In addition, reasoning can be the most difficult structural aspect of argumentation for teachers to successfully integrate into their classroom practice (McNeill, 2009). Consequently, teachers may need support around the structural aspects of argumentation, such as what counts as evidence and reasoning, as well as instructional strategies to support students with these elements.

By its very nature, scientific argumentation is a dialogic process in which individuals socially construct, critique, challenge and revise claims about the natural world (Berland & Reiser, 2011). Scientists do not work in isolation, but rather engage in argumentation within a community through discourse over time (Osborne, 2010). Engaging students in this practice requires supporting students in being enculturated into a community with particular norms that include persuading or convincing each other of their claims (Berland, 2011). The Next Generation Science Standards (NGSS) (Achieve, Inc., 2013) describe this social process in Appendix F (focused on the science and engineering practices): "Argumentation is the process by which evidence-based conclusions and solutions are reached…Scientists and engineers use argumentation to listen to, compare, and evaluate competing ideas and methods" (p. 29). Although this process is important for k-12 classrooms, teachers can have difficulty with the dialogic elements of argumentation. For example, when analyzing classroom discussions in both videos and written transcripts, they can have a hard time noticing these characteristics and instead may focus on superficial aspects such as "The teacher used encouraging words" (McNeill & Knight, 2013, p. 956). Furthermore, teachers can have challenges supporting this type of dialogic culture in their own classrooms, even when using curriculum that encourage these types of interactions (Alozie, Moje & Krajcik, 2010; McNeill & Pimentel, 2010). Consequently, teachers may need greater support around the dialogic aspects of argumentation.

### *Pedagogical Content Knowledge*

For teacher educators to develop more effective experiences and programs, they need a better understanding of teachers' PCK (Schneider & Plasman, 2011). PCK includes a variety of complex knowledge essential for effective teaching. Since its original conception, PCK has been defined, translated and extended in a variety of ways by science educators (Abell, 2007). Two elements of PCK included in Shulman's (1986) original conception are frequently discussed within the science education literature: 1) Knowledge of students' conceptions and 2) Knowledge of instructional strategies (Park & Oliver, 2008). Knowledge of students' conceptions includes an understanding of learning goals for students, areas of those learning goals that can be challenging, as well as the ability to identify and evaluate students' contributions in classroom practice (Shulman, 1986). Teachers' ability to make sense of students' responses is essential for planning future science instruction (Park & Oliver, 2008). After assessing students' strengths and weaknesses, teachers then need to use appropriate

instructional strategies to meet those students' needs. Knowledge of instructional strategies includes not only an understanding of what strategies to use with students, but also when those strategies would be most effective (Shulman, 1986).

The amorphic nature of PCK makes it challenging for scaffolding PCK through teacher education experiences and assessing teachers' PCK (Park & Oliver, 2008). Teachers' PCK can be implicit and highly contextualized making it difficult for teachers to express their ideas, resulting in a complex construct not easily assessed (Baxter & Lederman, 1999). Despite these challenges, an understanding of teachers' PCK can be important for the design of teacher education experiences (Schneider & Plasman, 2011). Consequently, our work focuses on developing a measure of teachers' PCK of argumentation that can provide valuable information for the design and revision of teacher education experiences.

## PCK Assessment Development Process

Our development and piloting process builds off the framework proposed by Hill and her colleagues (2004, 2008) focusing on conceptualizing the domain, developing assessment items and field testing assessment items. Their model "ultimately connects all three pieces of this work, tying conceptualization directly to the specification of items, and tying results from field tests back to strengths and weaknesses of the initial conceptualization (2008, p. 373). We utilized this framework in our own iterative design resulting in three different versions of the PCK of argumentation assessment, each of which was informed by the conceptualization and data from the previous steps. Table 1 provides a summary of our development process to date. Similar to Hill and her colleagues, we began with conceptualizing the target domain (Step 1), designing the first version of the items to align with those conceptions (Step 2) and then pilot testing the items (Step 3) as well as conducting cognitive interviews with teachers (Step 4). The data from the pilot testing and cognitive interviews was then used to revise the items resulting in the second version. We then added an additional component in that we asked ten external advisors to provide feedback on the items to assess the construct validity of the measure (Shadish, Cook & Campbell, 2002). Based on this feedback, we engaged in another round of revision developing a third version of the assessment. We will next discuss each step in more detail using sample items from one vignette to illustrate the development process.

***Table 1: Development Process for PCK of Argumentation Items***

| Step | Description of Step |
|---|---|
| 1. Conceptualization of the domain | • Conducted a literature review to develop initial 4 argumentation conceptions for PCK items |
| 2. Design of items (Version 1) | • Developed 8 vignettes each with 4 multiple-choice items and 1 open-ended item for a total of 32 multiple-choice and 8 open-ended items. |
| 3. Pilot testing of items | • Pilot tested 8 vignettes with 103 middle school teachers.<br>• Selected 6 vignettes for cognitive interviews based on greatest variation in teacher response. |
| 4. Cognitive interviews | • Conducted cognitive interviews with 24 middle school teachers for the 6 vignettes that remained after the pilot test cut. |
| 5. Revision #1: Items (Version 2) | • Revised all 6 vignettes using the data from both the pilot test and cognitive interviews. |

| | |
|---|---|
| | • For the pilot test data, examined the variation in teacher responses to the items. Found that distractors need to be appealing and not assessing surface level features.<br>• For the cognitive interviews, examined the teachers' rationales for selecting the choices, ideally wanting them to use PCK of argumentation and not other knowledge (i.e. test taking) to select the correct choice. |
| 6. Advisory board feedback | • Selected 4 vignettes to receive feedback from the advisory board based on two criteria: 1) Science content was more challenging for two of the vignettes (density and electromagnets), which appeared to shift the focus away from argumentation and 2) Included two vignettes where there was a correct claim and two vignettes where there was not a correct claim. Revised items to have even distribution for the 4 conceptions.<br>• Asked 10 advisors to provide the correct answer for each item, rate how well the item aligned with the conception and provide feedback for revision. |
| 7. Revision #2: Items (Version 3) | • Revised 4 vignettes based on whether the advisors selected the correct response, their ratings for the items and feedback. Revisions also took into consideration the teacher data from Revision #1 in order to not contradict any previous changes. |

***Step 1: Conceptualization of the domain.*** In Fall 2012, we began by conceptualizing the domain of argumentation by reviewing the discussion of the scientific practice in *A Framework for k-12 Science Education* (NRC, 2012), the draft version of *NGSS* (Achieve Inc., 2012) and relevant research. In reviewing the literature, we were interested in two areas. First, we were interested in identifying argumentation conceptions that previous research had found to be difficult for teachers (e.g. Crippen, 2012; Sampson & Blanchard, 2012; McNeill & Knight, 2013; Sadler, 2006; Zembal-Saul, 2009). We felt that these were areas of promise for supporting future teacher learning. Since the research focused on teachers' understandings of argumentation is fairly limited, we also examined the literature for areas of difficulty for students (e.g. Berland & Reiser, 2009; Osborne, Erduran & Simon, 2004; Sampson & Clark, 2011; Sandoval & Millwood, 2005). Our rationale was if there were areas in which students were having difficulty, teachers would also want greater support in these areas, again making them promising areas for teacher learning. In addition, we looked at our own work in terms of teachers' enactment of argumentation curriculum. We plan to use the PCK of argumentation assessment in conjunction with our own multimedia educative curriculum materials (MECM) that are being developed to support middle school science teachers in argumentation (Loper, McNeill, Peck, Price & Barber, 2014). Consequently, we wanted to target areas of argumentation that had proven to be challenging in previous teachers' enactment of the curriculum (McNeill, Gonzalez-Howard, Katsh-Singer, Price & Loper, 2013).

This definition process included delineation of boundaries to identify what goes beyond our designed measure (Hill et al., 2008). Instead of focusing on all aspects of argumentation, we decided to focus on a smaller number of areas to target the development of our measure. Ultimately, we focused on four conceptions related to argumentation with the first two conceptions focused on the structure of an argument and the last two conceptions focused on argumentation as a dialogic process (Table 2). The s*tructural* aspect focuses on argumentation as

a reasoned piece of discourse in which a claim is supported by an appropriate justification. Specifically, students need to support a claim with evidence and reasoning that uses scientific ideas to explain the link between the evidence and claim (McNeill et al., 2006). Conception 1 focuses on students using high quality evidence, such as measurements and observations, rather than low quality (i.e. data from am unreliable source) or non-evidence (i.e. students' opinions). Conception 2 emphasizes that students should articulate their reasoning, explaining why their evidence supports their claim using appropriate scientific ideas.

We also considered a *dialogic* aspect, which focuses on argumentation as persuasion or the interactions that occur between individuals when they try to convince an audience about the strength of a particular claim (McNeill & Pimentel, 2010; Jiménez-Aleixandre & Erduran, 2008). We decided to focus on two goals in relation to this dialogic aspect. Conception 3 focuses on the idea that there are multiple individuals interacting during an argument as they try to persuade each other of the strength of a claim. The process of attempting to persuade a community of the strength of a claim reveals the weaknesses in the argument and ultimately helps the community improve upon the ideas being discussed (Berland & Reiser, 2011). Conception 4 emphasizes the importance of considering multiple claims as part of the argumentation process. Considering multiple claims can support greater student understanding since knowing why an idea is wrong is as important as knowing why an idea is correct (Osborne, Simon, Christodoulou, Howell-Richardson & Richardson, 2013). Furthermore, without the inclusion of multiple potential claims, classroom instruction can end up focusing on the presentation of the scientifically accurate explanation rather than a process of critiquing and revising claims (Berland & McNeill, 2010).

***Table 2: Four Argumentation Conceptions for the Development of PCK Items***

| Conception | Title | Description |
|---|---|---|
| Conception 1: Evidence | Students use high-quality evidence to support their claims. | • Students understand the function of high-quality evidence in an argument, which includes making an argument stronger and more convincing<br>• High quality evidence consists of data, such as accurate measurements and observations, from a reliable source<br>• This conception promotes high-quality evidence over low-quality (i.e. data from an unreliable source) and non-evidence (students' opinions) |
| Conception 2: Reasoning | Students use scientific ideas or principles to explain the link between their evidence and claim (reasoning). | • Students understand that the use of scientific principles make the connection between the evidence and claim in an argument more clear<br>• This conception promotes students' use of clear and complete scientific reasoning over unclear, or absent reasoning. Clear and complete reasoning includes the logic behind why the evidence supports the claim using disciplinary core ideas. Unclear reasoning mentions the disciplinary core idea without much explanation OR does not articulate the logic behind why the disciplinary core idea is important |
| Conception 3: Persuasion | Students engage in argumentation with persuasion as the goal. | • Students consider the persuasion of an audience given that argumentation is a social process that includes multiple individuals interacting<br>• Persuasion includes critiquing and questioning arguments |

| | | |
|---|---|---|
| | | produced by others, as well as listening and building off of others' ideas |
| Conception 4: Multiple Claims | Students consider multiple claims as they engage in argumentation. | • Students consider and critique multiple claims as they work together as a community to address the specific question or problem<br>• Students understand that not only does considering multiple claims improve the quality of an argument but it also is an important practice for continually improving scientific explanations<br>• Students consider alternative or multiple claims when developing and/or articulating a rebuttal to their argument |

***Step 2: Design of items.*** We utilized the four argumentation conceptions to develop both multiple-choice and open-ended items targeting PCK of argumentation. PCK is highly contextualized in classroom practice (Park & Oliver, 2008; Hill et al., 2008). Consequently, we designed our assessment items to focus on vignettes illustrating strong and weak instances of argumentation in classroom discussion and student writing, rather than asking more general questions about argumentation (e.g. What is a scientific argument?). Each vignette focused on a fictional teacher's classroom. The vignettes begin by providing the lesson context such as the science topic and current question being investigated by the students. Then the vignettes include four multiple-choice items and one open-ended item that use samples of student writing and classroom talk to focus on students' conceptions of argumentation (e.g. Is the student struggling with some aspect of argumentation?) and instructional strategies (e.g. What would be an effective instructional strategy at this point during the lesson?). We purposefully designed each item to target one of the four conceptions. Initially, we developed eight vignettes for a total of thirty-two multiple-choice items and eight open-ended items. Version 1 of the PCK of Argumentation Assessment with all eight vignettes can be found in the Supplementary Materials (Methods S1). Table 3 provides a summary of the eight vignettes in Version 1 of the assessment.

***Table 3: Summary of Vignettes and Rationales for their Removal***

| **8 Initial Vignettes** | | | | **Selection of 6 Vignettes: Based on Teacher Pilot Data** | **Selection of 4 Vignettes: Based on Teacher Cognitive Interviews** |
|---|---|---|---|---|---|
| Fictional Teacher | Science Content | Question | Possible Claim(s) | | |
| Mr. Cedillo | Physical Science: Friction | Which type of material will allow a car to travel the fastest? | One | ✓ | ✓ |
| Ms. Moore | Physical Science: Density | Which ball(s) will sink? | One | ✓ | Removed – Vignette only had one correct claim. Density was a more challenging concept and influenced teachers' responses. |
| Mr. Luongo | Life Science: Characteristics of plants and animals | Should *elysia chlorotica*, a unique species of sea slug, be characterized as a plant or animal? | Multiple | ✓ | ✓ |
| Mr. Strong[1] | Life Science: Human needs for survival | Could humans survive in settlements on Mars? | Multiple | ✓ | ✓ |
| Ms. Salazar | Physical Science: Electromagnets | Which type of electromagnet is the strongest? | One | ✓ | Removed - Vignette only had one correct claim. Electromagnetism was a more challenging concept and influenced teachers' responses. |
| Ms. Alves | Earth Science: Plate tectonics | Have these two land masses always been in the same location? | One | ✓ | ✓ |
| Ms. Han | Life Science: Living things | Should viruses be classified as alive or not alive? | Multiple | Removed - Vignette with the least amount of variation in response. For 3 MC items, the majority answered the item correctly (82%, 82% and 76%) while the last item the majority answered incorrectly (22%). | Previously Removed |
| Mr. Lewis | Life Science: Biodiversity | Is the biodiversity in our schoolyard high or low? | Multiple | Removed – Vignette with the second lowest variation in response. For 3 MC items, the majority answered the item correctly (76%, 71% and 62%), while the last item had more variation (45%). | Previously Removed |

[1] Note: In the final version, we changed Mr. Strong to Ms. Strong so there would be 2 male and 2 female teachers.

Figure 1 includes one vignette to illustrate the format of the items. The two sample questions come from Mr. Cedillo's vignette in which the students investigated the question - Which type of material will allow a car to travel the fastest?

***Figure 1: Introduction, Question 1 and Question 2 from Version 1 - Mr. Cedillo Vignette***

Mr. Cedillo's students are analyzing the data table from an investigation they conducted that answered the question: Which type of material will allow a car to travel the fastest? The students timed how long it took for a toy car to travel 1 meter over a rug, wood floor, rubber mat, and ice.

| Surface | Distance Traveled (meters) | Time (seconds) |
|---|---|---|
| Rug | 1 | 10 |
| Wood floor | 1 | 5 |
| Rubber mat | 1 | 7.5 |
| Ice | 1 | 4 |

Ellen raises her hand in class and states the following argument: *The car on the ice will always go the fastest. I've been in a car driving on ice, and I know a car can skid because ice is the smoothest surface. My dad has a really big truck and it doesn't slide as far, so maybe next time we should try this experiment with larger cars.*

1. Mr. Cedillo should respond by saying:
   a. "Good job. Could someone else share a similar experience?"
   **b. "Great connection. Can anyone suggest data to support this?"**[1]
   c. "Nice argument. What additional evidence could Ellen add?"
   d. "Well done. Does anyone else want to present their argument?"

Mr. Cedillo next asks his students to engage in argumentation where they debate their ideas about the relationship between surface material and speed. The excerpt below is from the beginning of their conversation.

*Maya: My claim is that rough materials cause cars to go faster.*

*Elana: I think the data table shows that rough materials make cars go slower.*

*Ben: Well, I think there are lots of reasons a car would go faster or slower.*

2. Mr. Cedillo should speak up and encourage the students to:
   a. Raise their hands before sharing their ideas
   b. Focus on the scientifically accurate claim
   c. Review the vocabulary from the content wall
   **d. Persuade each other of the strength of their claim**[1]

[1] Correct answer choice is bolded.

The first two questions in Mr. Cedillos' vignette focus on a classroom discussion in which Ellen initially offers an idea, the focus of Question 1, and then three other students engage in the discussion. Question #1 focuses on Conception 1: Students use high quality evidence to support their claim. We hypothesized that in order to select the best response (choice b), teachers need to use PCK of argumentation, specifically around the quality of scientific evidence, to

analyze Ellen's contribution. A teacher would need to recognize that while Ellen provides an interesting everyday connection to her experiences in her father's car, she is not offering any scientific evidence or empirical data to support her claim. Question #2 focuses on Conception 3: Students engage in argumentation with persuasion as the goal. We hypothesized that to select the best response (choice d) requires teachers to use PCK of argumentation specifically around persuasion as a goal of argumentation to determine an appropriate instructional move for Mr. Cedillo. For Question #2, a teacher would need to recognize that the three students were presenting their claims, but not trying to convince each other of their claims.

Although we initially designed eight vignettes, our ultimate goal was to develop a final assessment that included four vignettes. We used the data from the seven steps of the design process to both revise items and remove vignettes. Table 3 provides a summary of the rationales for why 4 of the vignettes were removed. Our reasoning behind removing vignettes as well as revising the remaining vignettes will be discussed in more detail in the following steps.

***Step 3: Pilot testing items.*** After the design of the initial items, we pilot tested Version 1 with the 8 vignettes and associated PCK items with 103 middle school science teachers in spring 2013 using an online survey. Middle school science teachers were recruited through e-mail list servs in which they were invited to participate in a survey related to argumentation. Specifically, the e-mail stated, "For this project, we are hoping to get as many teachers as possible to complete a survey specifically related to argumentation …Your responses will be incredibly valuable in helping us improve the usefulness of the materials in supporting teachers." Consequently, the recruited teachers knew the survey focused on argumentation, but not that it focused on assessing PCK of argumentation. Upon completion of the survey, teachers received a $40 Amazon gift card. We originally intended to accept the first 100 teachers who completed the survey. However, multiple teachers were simultaneously completing the survey when we reached 100. Consequently, all of those teachers finished the survey resulting in 103 participants. Table 4 includes the background of the participants.

***Table 4: Teachers' Backgrounds for the Survey (n =103)***

| Type of Teaching Credential(s)[1] | None | Elementary | Middle or Secondary Science | TESOL, ESOL or ESL | SpEd | Other |
|---|---|---|---|---|---|---|
| *# of teachers* | *1* | *33* | *93* | *4* | *4* | *18* |
| Years of Teaching Experience | 1 | 2 – 5 | 6 – 10 | 11- 15 | 16 – 20 | > 20 |
| *# of teachers* | *2* | *6* | *36* | *23* | *12* | *24* |
| Highest Degree in Education | None | Bachelors | Masters | Doctorate | | |
| *# of teachers* | *2* | *29* | *68* | *4* | | |
| Highest Degree in Science | None | Bachelors | Masters | Doctorate | | |
| *# of teachers* | *21* | *56* | *26* | *0* | | |
| Argumentation Workshops | 0 | 1 | 2 or 3 | 4 or more | | |

| | | | | |
|---|---|---|---|---|
| Attended | | | | |
| *# of teachers* | *45* | *24* | *21* | *13* |
| Use of Argumentation in Classroom | Never | Once | A few times | Many times |
| *# of teachers* | *19* | *7* | *53* | *24* |

[1] Teachers could provide multiple answers for "Type of Teaching Credential(s)"

Because of the length of the assessment and our concern that teachers would not complete it, we split the assessment into two versions each containing four vignettes. Each teacher completed one version of the assessment with 48 teachers completing PCK Assessment A (PCK A) and 55 teachers completing PCK Assessment B (PCK B). We used the results of the survey to inform our revision process. Table 7 provides a summary of the descriptives for the multiple-choice items for the two versions of the assessment. For each conception, we saw that on average the teachers selected the correct answer about half of the time with PCK A potentially being a little easier than PCK B. The teachers taking the survey had a range of backgrounds and experience with argumentation (see Table 4) so overall we felt this was appropriate. If anything, we wanted the assessment to be even harder in order to enable greater sensitivity to teacher growth in relation to professional development or educative curriculum experiences. To check the reliability of the assessment, Cronbach's alpha was calculated separately for the two assessments. PCK A's Cronbach's alpha was 0.222 and PCK B's Cronbach's alpha was 0.500. We were not surprised that these reliabilities were low considering this was our initial attempt at developing such an assessment and we expected that significant revision would need to occur.

***Table 5: Descriptives for Multiple-Choice Items***

| | PCK A (n = 48)[1] Mean, SD (Max) | PCK B (n = 55)[2] Mean, SD (Max) |
|---|---|---|
| Conception 1: Evidence | 2.77, 1.12 (max = 5) | 2.82, 1.00 (max = 4) |
| Conception 2: Reasoning | 2.38, 1.00 (max = 4) | 1.73, 0.87 (max = 3) |
| Conception 3: Persuasion | 1.88, 0.87 (max = 3) | 2.00, 0.84 (max = 5) |
| Conception 4: Multiple Claims | 1.96, 1.03 (max = 4) | 1.89, 1.15 (max = 4) |
| Total Multiple-Choice | 8.98, 2.19 (max = 16) | 8.44, 2.43 (max = 16) |

[1] PCK A consisted of the Mr. Cedillo, Ms. Moore, Mr. Luongo, and Mr. Strong vignettes
[2] PCK B consisted of the Ms. Salazar, Ms. Alves, Ms. Han, and Mr. Lewis vignettes

Next, we used the item level descriptive statistics (i.e. percentage of respondents for each choice) to select six of the eight vignettes as the focus of our continued development work. We determined the descriptive statistics to examine the spread across the four choices for each multiple-choice item (See Supplementary Materials – Table_S1). We selected six of the eight vignettes for the cognitive interviews, because they included the greatest variation in teacher responses based on the descriptive statistics (see Table 3). The two vignettes that we eliminated, Ms. Han and Mr. Lewis, included the majority of teachers providing the "right" answer. Considering that many of the teachers in the sample had little experience with argumentation (see Table 4), we felt that these items were too easy for their intended purpose. Specifically for

Ms. Han, for three of the multiple-choice items the majority of teachers responded correctly (82%, 82% and 76%), while the last item the majority answered incorrectly (22%). Similarly for Mr. Lewis, for three of the multiple-choice items the majority of teachers responded correctly (76%, 71% and 62%) while the last item had more variation (45%). The lack of variation in teachers' responses for these vignettes would not enable us to distinguish the knowledge levels between teachers or offer room for growth when using the items in association with teacher education experiences. Consequently, we decided to remove these two vignettes and focus our revision efforts on the remaining six vignettes. We will return to this data in Step 5 where we will discuss how we used them to revise the remaining items.

***Step 4: Cognitive interviews.*** We conducted cognitive interviews with 24 middle school science teachers about their responses to the six remaining vignettes from Version 1 of the PCK of argument assessment. Cognitive interviews can be used to learn about participants' thought processes as they construct responses to selected assessment items (Wilson, 2005). Specifically, when designing a teacher assessment, cognitive interviews provide an important measure of validity to determine whether teachers are using the targeted understandings and not other knowledge such as test-taking strategies (Hill et al., 2008).

We recruited middle school science teachers by e-mailing local teachers who had previously participated in at least one professional development workshop about argumentation with the first author. We targeted local teachers, because we wanted to conduct the cognitive interviews in person so that if during the interview the teacher pointed to or indicated a section of the assessment, we could follow-up with appropriate questions. We felt these visual cues would be important to understanding how they were responding to the items. Furthermore, we decided to focus on teachers who had some experience with argumentation to see if they utilized that knowledge in answering the questions. We interviewed the first 24 teachers who responded to the e-mail. Upon completion of the interview, the participants received a $75 Amazon gift card.

Table 6 includes a summary of the middle school teachers' backgrounds. Although they include a range of backgrounds, the minimum number of years teaching was 2 years with many of the teachers having been in the classroom for numerous years. All of the teachers had attended at least one argumentation workshop and reported integrating argumentation into their classroom at least a few times.

***Table 6: Teachers' Backgrounds for the Cognitive Interviews (n = 23)[1]***

| Type of Teaching Credential(s)[2] | Elementary | Middle or Secondary Science | TESOL, ESOL or ESL | SpEd | Middle School Math | |
|---|---|---|---|---|---|---|
| *# of teachers* | *3* | *21* | *3* | *5* | *3* | |
| Years of Teaching Experience | 1 | 2 – 5 | 6 – 10 | 11- 15 | 16 – 20 | > 20 |
| *# of teachers* | *0* | *5* | *6* | 6 | *3* | *3* |
| Highest Degree in Education | None | Bachelors | Masters | Doctorate | | |
| *# of teachers* | *3* | *0* | *20* | *0* | | |
| Highest Degree in Science | None | Bachelors | Masters | Doctorate | | |

| | | | | |
|---|---|---|---|---|
| *# of teachers* | *6* | *12* | *5* | *0* |
| Argumentation Workshops Attended | 0 | 1 | 2 or 3 | 4 or more |
| *# of teachers* | *0* | *10* | *8* | *5* |
| Use of Argumentation in Classroom | Never | Once | A few times | Many times |
| *# of teachers* | *0* | *0* | *10* | *13* |

[1] Although 24 cognitive interviews were conducted, one teacher did not provide background information.
[2] Teachers could provide multiple answers for "Type of Teaching Credential(s)"

Because of the length of the assessment, we had each teacher respond to only three vignettes, which resulted in a total of twelve multiple-choice items and three open-ended items. The teachers were asked to write their responses to the assessment and to "think-aloud as you write your responses." If the teachers were silent for awhile, the interviewer would follow-up with a prompt such as, "What are you thinking?" or "Can you continue speaking?". The interviews ranged from about 25 minutes to 60 minutes, with the majority lasting approximately 40 minutes. In the next step, we discuss how we used the data from the cognitive interviews in the revision of the items.

***Step 5: Revision of the items.*** Next, we used the results from both the pilot test and the cognitive interviews to revise the assessment items for the six vignettes in Version 1 of the Assessment. For the six open-ended items, we developed coding schemes to analyze the teachers' responses for both the surveys (i.e. Step 3) and the cognitive interviews (i.e. Step 4). One rater coded each of the teachers' open-ended responses. We then randomly sampled 20% of the teachers, which were scored by a second independent rater. Our estimates of inter-rater reliability were calculated by percent agreements. Our inter-rater agreement for the six-open ended items was 93% for the survey responses and 91% for the cognitive interview responses.

In addition, for the cognitive interviews, we developed a coding scheme to capture the teachers' rationales for their responses to the assessment items. In Hill and her colleagues' work (2008), they coded teachers' responses for knowledge of content and students (similar to PCK), mathematical reasoning and test-taking skills. We built on this coding scheme, but expanded it to include six categories of responses: 1. Accurate PCK of argument, 2. Inaccurate PCK of argument, 3. Science content, 4. Literacy, 5. Test-taking skills, and 6. Other rationale, unclear rationale or no rationale (see Table 7). One rater coded each teacher's interview. We then randomly sampled 25% of the teachers' interviews, which were scored by a second independent rater. Our estimates of inter-rater reliability were calculated by percent agreements. Inter-rater agreement was 83%.

Table 7 provides a description of each of the codes for the teacher rationales as well as shows the frequency and number of responses that each rationale was used across all of the items. The teachers received multiple codes for each multiple-choice item since we coded for each answer choice (i.e. why they selected "a" as well as why they did not select "b", "c" and "d"). Overall, we saw that Code 1: Accurate PCK of argument was the most frequent rationale for answering the items. If individuals were experts in argumentation and the instrument was

valid, we would hope to see Code 1 used 100% of the time. Although we purposefully interviewed teachers with experience with argumentation, not surprisingly the group included a range of expertise. Consequently, in terms of revising the assessment, we still viewed responses coded as Code 2: Inaccurate PCK of Argument, as positive, since the assessment was targeting PCK of argument. When teachers used other rationales (i.e. Codes 3-6), we then considered how to revise the items to target PCK of argument. To do this, we created item level descriptives for each question to examine what type of rationales the teachers were using in selecting their answer choice (See supplementary materials Table S2). We will next illustrate how we used this item level information in combination with the item specifics from the pilot data (Table S1) to revise each item in the 6 vignettes.

***Table 7: Teachers Rationales from the Cognitive Interviews (n=24)***

| Code | Description | Use of Rationale[1] |
|---|---|---|
| Code 1: Accurate PCK of Argument | Teacher uses accurate knowledge of students' conceptions or strategies for teaching argumentation to select the response | 49.3% (603) |
| Code 2: Inaccurate PCK of Argument | Teacher uses inaccurate knowledge of students' conceptions or strategies for teaching argumentation to select the response | 19.8% (242) |
| Code 3: Science Content | Teacher uses knowledge of the science content (e.g. density) to select the response | 4.7% (58) |
| Code 4: Literacy | Teacher discusses general strategies for literacy, but are not specific for argument, for selecting the response | 0.1% (1) |
| Code 5: Test Taking Skills | Teacher uses information from the stem to match to the choice or eliminates choices to get to the final response | 4.1% (50) |
| Code 6: Other rationale, Unclear rationale or No Response | Teacher provides a rationale that does not align with previous codes, provides an unclear rationale or does not provide a response. | 22.1%(270) |

[1] This includes the percentage and total number of rationales given each code.

To illustrate this process, we return to Question #1 from the Mr. Cedillo vignette presented in Figure 1, which asked teachers how Mr. Cedillo should respond to Ellen's comment that her dad's really big truck does not slide as far on ice. Table 8 provides a summary of the number of teachers who selected each answer choice for Question 1. Table 9 includes the percentage of teachers we coded for each category of rationale and sample quotes from the cognitive interviews. For Question 1, the least common answer was "choice a" suggesting that the teachers did not see it as the most productive teacher response to support the argumentation lesson. For example, Teacher 006 explained that she did not pick choice a, because, "I think that 'a' is just acknowledging, it's kind of a thank you for sharing." We coded this rationale as Code 6, because her response focused on the fact that the comment was general and not necessarily argument specific. Only one of the sixty teachers across both the cognitive interviews and survey selected "choice a" suggesting that it was easy for all of the teachers to rule out regardless of their understanding of argument. Consequently, the data suggested that we should revise "choice a" to make it a more appealing distractor.

*Table 8: Teacher Choices for Version 1 Mr. Cedillo Questions 1*

| Answer Choice[1] | Cognitive Interview (n = 12) | Survey (n = 48) |
|---|---|---|
| a. "Good job. Could someone else share a similar experience?" | 8% (1) | 0% (0) |
| **b. "Great connection. Can anyone suggest data to support this?"** | **25% (3)** | **48% (23)** |
| c. "Nice argument. What additional evidence could Ellen add?" | 58% (7) | 37% (18) |
| d. "Well done. Does anyone else want to present their argument?" | 8% (1) | 15% (7) |

[1] Bold choice is the correct answer.

Choices b, c and d were more frequently selected. All three of these choices included terms related to argument (i.e. data, evidence, and argument) suggesting that argumentation specific answer choices were more appealing to teachers. We see from the teachers' rationales for selecting these choices that teachers had different understandings of what counted as evidence for an argument, which impacted their answer choice selection. For example, Teacher 012 who correctly selected "choice b" used PCK that aligned with Conception 1 about high quality evidence, which was the focus of this item, so her rationale was given Code 1. Specifically, the teacher articulated that Ellen "should actually be using the experiment" for her evidence rather than a personal story. Teachers who selected choices c and d often failed to recognize that Ellen's response does not include empirical data. For example, Teacher 018 who selected "choice c" stated, "she's got some evidence there". We gave her rationale a Code 2 for inaccurate PCK of argument. While she was drawing on knowledge of argumentation to select her choice rather than other knowledge like test-taking abilities, that knowledge was incorrect.

*Table 9: Teacher Rationale for Version 1 Mr. Cedillo Question #1 (n = 12)*

| Choice[1] | Rationale | Sample Teacher Response for Most Frequent Rationale |
|---|---|---|
| a. 8% | Code 1 – 42% (5)<br>Code 3 – 17% (2)<br>Code 5 – 8% (1)<br>Code 6 – 33% (4) | Teacher 006 decided not to select choice "a", because "I think that 'a' is just acknowledging, it's kind of a thank you for sharing." (Code 6) |
| **b. 25%** | Code 1 – 17% (2)<br>Code 2 – 58% (7)<br>Code 5 – 8% (1)<br>Code 6 - 17% (2) | Teacher 012 selected choice "b", because "…ultimately we want to be using this evidence to state their argument rather than just their background knowledge. This just is a great story, but we should actually be using the experiment." (Code 1) |
| c. 58% | Code 1 – 42% (5)<br>Code 2 – 42% (5)<br>Code 6 - 17% (2) | Teacher 018 selected choice "c" because "She's making an argument, and she's got some evidence there but we want to back it up with more evidence." (Code 2) |
| d. 8% | Code 1 – 33% (4)<br>Code 2 – 17% (2)<br>Code 3 - 8% (1)<br>Code 6 – 42% (5) | Teacher 024 selected choice "d" because "I think she gave the data, so I'm going to go with D as the best response." (Code 2) |

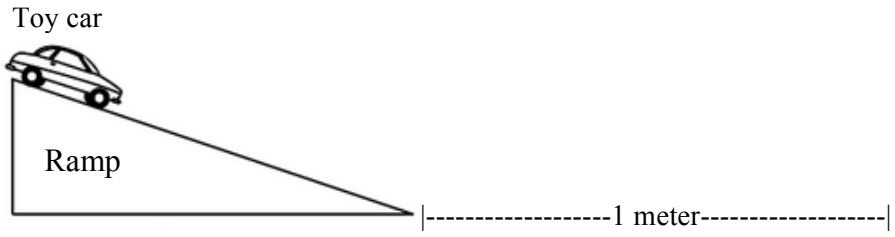
[1] Bold choice is the correct answer.

Since teachers typically used either accurate (Code 1) or inaccurate (Code 2) PCK of argument for answer choices b and c, and these two choices frequently had a number of teachers selecting them, we did not revise these choices. Answer choice d also appeared to encourage teachers to use accurate or inaccurate knowledge of PCK; however, this answer choice was selected less frequently (i.e. 8% for interview and 15% for survey). Consequently, we decided to make a minor revision to the item. Specifically, during the cognitive interviews teachers were more likely to talk about wanting their students to "share" their ideas rather than "present" their

ideas. The term "sharing" appeared to be more appealing perhaps because it suggests more collaboration or a classroom community. Consequently, we revised this term. Furthermore, teachers raised questions during the cognitive interviews about the set-up of the experiment in the vignette, and so revisions were made to provide more experimental context. Figure 2 includes the revised opening and Question 1 from the Mr. Cedillo vignette that was included in Version 2 of the instrument, the version that was shared with the external advisors (step 6).

***Figure 2: Introduction and Question 1 from Version 2 - Mr. Cedillo Vignette***

Mr. Cedillo's 7th grade science class is doing a unit on force and motion. Near the middle of the unit his students are exploring friction by analyzing the data table from an investigation they conducted that answered the question: Which type of surface material will allow a toy car to have the greatest average speed? The students let a toy car go from the top of a ramp and timed how long it took to travel 1 meter after reaching the bottom of the ramp, over four different surface materials: a rug, wood floor, rubber mat, and ice (see image below).




They then calculated the toy car's average speed by dividing the distance over the time. The table below shows the students' experimental results.

| Surface Material | Distance Traveled (meters) | Time (seconds) | Average Speed (meters/seconds) |
|---|---|---|---|
| Rug | 1 | 10 | 0.10 |
| Wood floor | 1 | 5 | 0.20 |
| Rubber mat | 1 | 7.5 | 0.13 |
| Ice | 1 | 4 | 0.25 |

Ellen raises her hand in class and states the following argument: The car on the ice will always go the fastest. I've been in a car driving on ice, and I know a car can skid because ice is the smoothest surface. My dad has a really big truck and it doesn't slide as far, so maybe next time we should try this experiment with larger cars.

1. Mr. Cedillo should respond by saying:
   a. "Interesting point, Ellen. Does anyone have similar reasoning?"
   **b. "Great connection. Can anyone suggest data to support this?"[1]**
   c. "Nice argument. What additional evidence could Ellen add?"
   d. "Well done. Does anyone else want to share their argument?"

[1] Correct answer choice is bolded.

We engaged in a similar process for all six vignettes resulting in the revision of twenty-four multiple-choice items and six open-ended items. Our overarching goal for this revision was

to develop items that explicitly targeted the four conceptions (see Table 2) and distinguished between the knowledge levels of teachers.

***Step 6: Advisory board feedback.*** To assess the construct validity of our PCK of argumentation assessment, we asked for external feedback from ten advisors. We shared the items with the advisors and had them evaluate whether each item aligned with the theorized construct (Shadish, Cook & Campbell, 2002). We selected the advisors based on their expertise in various aspects of argumentation (i.e. writing, assessment, etc.). All ten advisors held doctorate degrees and have published peer reviewed journal articles about argumentation or scientific practices. Nine of the advisors were faculty members at universities and one of the advisors worked at an education non-profit organization. The advisors were provided an honorarium to compensate them for their time.

Specifically, the advisors were e-mailed a document summarizing the four argumentation conceptions (Table 2) and an online survey which included 4 vignettes from the PCK of argumentation Version 2 items (see Supplementary Materials - Methods_S1). Before contacting the advisors, we eliminated two more vignettes, Ms. Moore and Ms. Salazar, because we felt the length of the assessment was still too long for someone to complete in one sitting. We removed these two vignettes based on two criteria. First, we decided the final assessment should include two vignettes with one possible correct claim and two vignettes with multiple possible correct claims. Some of the vignettes were designed such that there was only one possible claim that students could correctly support with evidence, like Mr. Cedillo's vignette about friction. Other vignettes had more than one possible claim, like Mr. Strong's vignette, which we will discuss in more detail later that addresses the question – Could humans survive in settlements on Mars? (see Figure 4). Since we decided we wanted an even distribution, this meant we needed to remove two vignettes that only had one claim. When we discussed which two of the four vignettes with one claim to remove, we realized from the cognitive interviews that the science content in two of the vignettes was more challenging for some teachers. Specifically, Ms. Moore focused on density and Ms. Salazar focused on electromagnets. Although we attempted to remove any issues with the content in our revisions, we were still concerned that teachers' knowledge of the science content could be impacting their responses rather than their PCK of argument. Consequently, we removed these two vignettes. This left the final four vignettes which are described in Table 3.

The advisors were asked to answer each multiple-choice and open-ended item in the assessment. Our rationale was that if these ten advisors could not correctly answer the items then the items may not be measuring PCK of argumentation, and so we would consider these items for revision (step 7). Furthermore, for each assessment item we asked the advisors to rate the quality of the item. Specifically, they were asked, "How well do you think this question aligns with Conception [1, 2, 3 or 4]?"[1] We filled in the appropriate conception for each question. They were then provided with the following likert choices: 1. Aligns Very Well, 2. Aligns, 3. Somewhat Aligns, and 4. Does Not Align. Next, they were provided with the following question, "Please provide any specific feedback for how this item can be improved to better assess PCK for Conception [1, 2, 3, or 4]." In addition, at the end of the assessment they were asked three overarching questions about the greatest strengths of the assessment, the greatest weaknesses of the assessment, and for any other additional feedback.

[1] Unfortunately, there were technical issues for two of the items and we did not receive likert ratings; however, we did receive written feedback that we used in the revision process.

***Step 7: Revision of items.*** After we received the advisors' feedback, we developed a set of rules to determine which items to revise. Table 10 includes a summary of the four rules as well as the number of items that aligned with each rule. We used the average rating of 2.0 and lower for the quality of the item, because 2.0 meant the item aligned well and 1.0 meant the item aligned very well. Consequently, if an item received a rating of 2.0 or lower, that suggested that on average the reviewers felt it aligned well or very well. Furthermore, ideally all of the reviewers would answer each item correctly. Consequently, if that was not the case, we wanted to consider revising the item.

***Table 10: Rules for Revision Based on Advisors Feedback***

| Rule | Description | # Items[1] |
|---|---|---|
| Rule 1 | If the item received an average rating of 2.0 or lower and all of the reviewers selected the correct answer, do NOT revise. | 6 |
| Rule 2 | If the item received an average rating of 2.0 or lower and one or more reviewers selected the incorrect answer, CONSIDER revising**.** | 6 |
| Rule 3 | If the item received an average rating higher than 2.0 and all of the reviewers selected the correct answer, CONSIDER revising. | 2 |
| Rule 4 | If the item received an average rating higher than 2.0 and one or more reviewers selected the incorrect answer, DEFINITELY revise. | 4 |

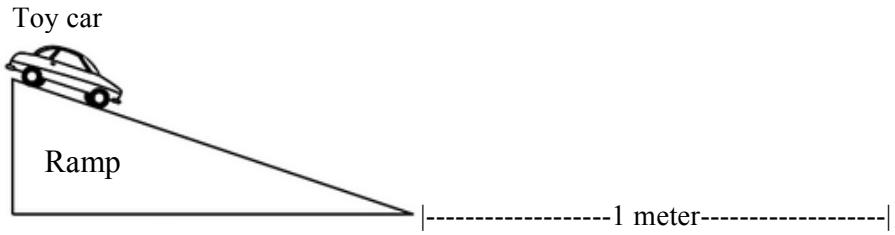
[1] This does not include the two items that did not receive ratings because of technical issues.

In order to illustrate how we used the advisors' feedback to revise individual items, we return to the vignette with Mr. Cedillo, specifically Question 1. The version of the item that advisors reviewed is in Figure 2. This item fell under the category of Rule 2, because all of the advisors answered the item correctly; however, it received an average rating of 2.1, which is between Aligns (2.0) and Somewhat Aligns (3.0). We read through the advisors' written feedback to better understand their concerns with the item. This resulted in the identification of two issues. The first issue was a concern that the car's times in the introduction were inaccurate. For example, Advisor 01 wrote, "…the times are unreasonably long and this will be a distraction for any teacher who has done this in class." Consequently, to address this issue we conducted the experiment to determine more accurate times (see Figure 3 for final item).

The second issue was that a couple of advisors were concerned that the distinction between the correct answer choice, b, and answer choice c was not explicit enough. For example, Advisor 2 wrote that to a teacher b and c "may both look appropriate, as the difference is a subtle one". Although we agree that the distinction is subtle, we felt this was an important distinction for teachers to learn about and would help differentiate the quality of teachers' PCK of argumentation. As we discussed in the cognitive interviews, many of the teachers selected c and d because they had different understandings of what counted as high quality scientific evidence (Table 9). In our own work designing teacher education experiences, we feel that this is something we would want to know either as part of a pre-assessment or as part of a post assessment if we were not able to help them develop a richer understanding through professional development or educative curriculum materials. Furthermore, all ten of the advisors did answer this item correctly, suggesting that an expert in the field could tell the difference. Consequently, we decided not to revise the answer choices.

***Figure 3: Introduction and Question 1 from Version 3 - Mr. Cedillo Vignette***

Mr. Cedillo's 7th grade science class is doing a unit on force and motion. Near the middle of the unit his students are exploring friction by analyzing the data table from an investigation they conducted that answered the question: Which type of surface material will allow a toy car to have the greatest average speed? The students let a toy car go from the top of a ramp and timed how long it took to travel 1 meter after reaching the bottom of the ramp, over four different surface materials: felt, top of lab table, sand paper, and ice (see image below).




They then calculated the toy car's average speed by dividing the distance over the time. The table below shows the students' experimental results.

| Surface Material | Distance Traveled (meters) | Time (seconds) | Average Speed (meters/seconds) |
|---|---|---|---|
| Felt | 1.0 | 2.4 | 0.42 |
| Top of lab table | 1.0 | 1.5 | 0.67 |
| Sand paper | 1.0 | 2.2 | 0.45 |
| Ice | 1.0 | 1.0 | 1.0 |

Ellen raises her hand in class and states the following argument: The car on the ice will always go the fastest. I've been in a car driving on ice, and I know a car can skid because ice is the smoothest surface. My dad has a really big truck and it doesn't slide as far, so maybe next time we should try this experiment with larger cars.

1. Mr. Cedillo should respond by saying:
    a. "Interesting point, Ellen. Does anyone have similar reasoning?"
    **b. "Great connection. Can anyone suggest data to support this?"[1]**
    c. "Nice argument. What additional evidence could Ellen add?"
    d. "Well done. Does anyone else want to share their argument?"

[1] Correct answer choice is bolded.

## Lessons Learned

In this section, we present four key lessons that we learned from the PCK of argumentation assessment development process. Table 11 provides a summary of these four lessons learned. We use examples from the data collected throughout this process (i.e. teacher pilot data, teacher cognitive interviews, and advisor feedback) to illustrate these lessons. Although the lessons stem from our development of a PCK of argumentation assessment, we feel the first three also have implications for the design of assessments targeting the PCK of other scientific practices. The last lesson learned is specific to argumentation.

***Table 11: Four Lessons Learned from PCK of Argumentation Assessment Development***

| | |
|---|---|
| 1 | In designing multiple-choice items, all distractors should focus on the targeted scientific practice and not other areas of science instruction. |
| 2 | In designing multiple-choice items, it is challenging to craft answer choices that assess a deep understanding of the scientific practice (rather than surface level features), but still have a clear correct answer. |
| 3 | Using vignettes is both a strength and weakness in the design of the items; although, the real world context is more authentic, this complexity also makes it more challenging to target the construct of interest. |
| 4 | For the two dialogic argumentation conceptions, it was more challenging to develop high quality items and to distinguish between the two conceptions. |

***Lesson #1: In designing multiple-choice items, all distractors should focus on the targeted science practice and not other areas of science instruction.***

In Version 1 of the assessment, we included distractors that focused on other aspects of science instruction, which we thought might be appealing for teachers who lacked PCK of argumentation. However, this did not turn out to be the case. For example, in our previous discussion of Mr. Cedillo Question 1, we found that only 1 of the 60 teachers across the cognitive interviews and survey, selected "choice a" that included a more generally worded distractor (i.e. Good Job. Could someone else share a similar experience?). Instead, teachers were more likely to select answer choices that were argument specific. This frequently occurred across the items in the assessment.

Mr. Cedillo Question 2 also illustrates this trend. In Version 1, the second question of this vignette included a sample transcript in which three students were talking, but not listening to or critiquing each other's ideas (Figure 1). The prompt then asked, "Mr. Cedillo should speak up and encourage the students to…". Table 12 includes the teachers' responses for this item.

*Table 12: Teacher Choices for Version 1 Mr. Cedillo Question 2*

| Answer Choice[1] | Cognitive Interview (n = 12) | Survey (n = 48) |
|---|---|---|
| a. Raise their hands before sharing their ideas | 8% (1) | 2% (1) |
| b. Focus on the scientifically accurate claim | 8% (1) | 17% (8) |
| c. Review the vocabulary from the content wall | 0% (0) | 0% (0) |
| **d. Persuade each other of the strength of their claim** | **83% (10)** | **81% (39)** |

[1] Bold choice is the correct answer.

In this item, choices a and c do not include argument specific ideas. None of the teachers selected choice c, while only 2 of the 60 teachers across the cognitive interview and survey selected choice a. Table 13 includes teachers' rationales for their choices from the cognitive interviews. The teachers' responses suggested that the teachers were not distracted by vocabulary-related answer choices (e.g. "reviewing the content vocabulary…it's connected but unrelated") or procedural related answer choices like students raising their hands (e.g. "hopefully that's already a procedure in place"). Consequently, this item was not effective at differentiating between the knowledge levels of the teachers participating in the pilot.

*Table 13: Teacher Rationale for Version 1 Mr. Cedillo Question #2 (n = 12)*

| Choice[1] | Rationale | Sample Teacher Response for Most Frequent Rationale |
|---|---|---|
| a. 8% | Code 1 – 75% (9)<br>Code 2 – 8% (1)<br>Code 5 – 8% (1)<br>Code 6 – 8% (1) | Teacher 006 decided not to select choice "a", because "Well, hopefully that's already a procedure in place. (Code 6) |
| b. 8% | Code 1 – 42% (5)<br>Code 2 – 25% (3)<br>Code 5 – 8% (1)<br>Code 6 – 25% (3) | Teacher 024 decided not to select choice "b", because "I don't know that they know what's the scientifically accurate claim." (Code 6) |
| c. 0% | Code 1 – 33% (4)<br>Code 2 – 8% (1)<br>Code 3 – 8% (1)<br>Code 5 – 8% (1)<br>Code 6 – 42% (5) | Teacher 020 decided not to select choice "c" because "And reviewing the content vocabulary, well I think that's important but I think that is, it's connected but unrelated at the moment to providing a claim for, you know, evidence for their claim." (Code 1) |
| **d. 83%** | Code 1 – 50% (6)<br>Code 2 – 42% (5)<br>Code 6 – 8% (1) | Teacher 014 selected choice "d" because "I would go with persuade each other because you're asking the kids to go back and evaluate their own claim and back it up with evidence." (Code 1) |

[1] Bold choice is the correct answer.

We revised the language of all non-argument distractors in this item and across the entire assessment before sharing Version 2 of the assessment with the advisors. We tried to focus all of the multiple-choice options on argumentation, either in terms of the structure of an argument or argumentation as a dialogic process.

***Lesson #2: In designing multiple-choice items, it is challenging to craft answer choices that assess a deep understanding of the scientific practice (rather than surface level features), but still have a clear correct answer.***

Even when all of the answer choices focused on argumentation, we still found there to be a challenge around having the items target a deep understanding of this scientific practice, rather than surface level features. Figure 4 includes Question 3 from the Mr. Luongo Vignette in which students investigated whether *elysia cholortica*, a unique species of sea slug, should be characterized as a plant or animal (see Methods S1 for complete vignette). We included in Figure 4 both Version 1, which the pilot teachers received, and Version 2, which the advisors received.

***Figure 3: Question 3 from Version 1 and Version 2 - Mr. Luongo Vignette***

Mr. Luongo then pairs up students to edit each other's arguments. While walking around the room he hears the following interaction:

**Leah:** Claire, you wrote that this slug becomes a plant after eating algae? You're using X-men to support your claim?

**Claire:** Yeah! Remember the character Rogue? She takes other mutants' powers and this slug basically does the same with algae—after eating algae it can do photosynthesis. So like Rogue this slug becomes what it takes in, in this case a plant.

**Leah:** Oh I guess you're right. I should add that as more supporting evidence for my claim too!

VERSION #1

3. After hearing these students' conversation Mr. Luongo should:
   a. **Have students review the concept wall's definition of what evidence is**[1]
   b. Encourage students to incorporate more everyday examples of evidence
   c. Remind students to include as many pieces of evidence as possible
   d. Ask students to use a graphic organizer to keep track of their evidence

VERSION #2[2]

3. After hearing these students' conversation Mr. Luongo should:
   a. **Ask students to review the concept wall's explanation of what evidence is**
   b. Encourage students to explain the scientific reasoning behind this evidence
   c. Remind students to incorporate as many pieces of evidence as possible
   d. Have students consider how this evidence could support the counter claim

[1] Correct answer choice is bolded.
[2] Underlined text was changed in Version 2.

Although all four answer choices in Version 1 included the word "evidence", the majority of teachers selected the correct answer, choice a. Table 14 includes the teachers' answer choices.

*Table 14: Teacher Choices for Version 1 Mr. Luongo Question 3*

| Answer Choice[1] | Cognitive Interview (n = 12) | Survey (n = 48) |
|---|---|---|
| **a. Have students review the concept wall's definition of what evidence is** | **92% (11)** | **71% (34)** |
| b. Encourage students to incorporate more everyday examples of evidence | 0% (0) | 8% (4) |
| c. Remind students to include as many pieces of evidence as possible | 8% (1) | 19% (9) |
| d. Ask students to use a graphic organizer to keep track of their evidence | 0% (0) | 1% (1) |

[1] Bold choice is the correct answer.

In discussing their rationales for their selections during the cognitive interviews, the teachers appeared to easily rule out "choice b", because, as Teacher 011 stated, "they were talking about

Hollywood" and Teacher 012 explained, "that is a good connection, but not evidence." Similarly, all of the teachers in the cognitive interview ruled out "choice d" using rationales such as Teacher 018, who stated, "It doesn't matter what you organize if you don't know what evidence is." Because these two choices were unappealing to the teachers, we completely revised them for Version 2. Answer choice c was interesting in that only one teacher during the cognitive interviews selected this option, while more teachers selected it during the survey. We felt this might be in part, because the teachers who participated in the cognitive interviews had more experience with argumentation. Consequently, we only made a minor change to this choice (i.e. switching "include" to "incorporate" to make the distractor a little longer to align with the others), because we felt it could potentially distinguish teachers with little PCK of argumentation.

After revising Mr. Luongo Question 3, we then provided Version 2 to our advisors for feedback. Overall, the advisors rated this item in between "aligns very well" and "aligns" with an average rating of 1.8. Some of the advisors felt this item did target a deep understanding of the practice. For example, Advisor 05 wrote, "The distractors in this item really help distinguish between possible understandings of what evidence is and/or how to support its use… this one is about ways of supporting evidence so gets more in-depth." Although the item now appeared to target a more in-depth understanding, two of the advisors selected the incorrect choice b. This suggests there is a tension in developing an in-depth item, but still having a clear correct answer. As Advisor 10 wrote, "I was actually torn between responses 1 and 2." A couple of the advisors, including those who did choose a, expressed concern with how "choice a" was currently worded, suggesting that students should just be using a definition of evidence that was provided to them. For example, Advisor 03 wrote "This worries me. It feels like orthodoxy. It isn't evidence because the definition on the board says it isn't evidence." and Advisor 07 wrote, "I have to confess I'm not sure an official definition of evidence on the classroom wall is in the spirit of argumentation." Consequently, in Version 3 of the assessment, we revised answer "choice a" to suggest that the definition was not imposed on the students, but instead was generated by the class. Specifically, we changed the wording to, "Prompt students to review the class description of what counts as evidence." Our revision process suggests that it is challenging to design answer choices that assess a deep understanding of the scientific practice (rather than surface level features), but still have a clear correct answer.

***Lesson #3: Using vignettes is both a strength and weakness in the design of the items; although, the real world context is more authentic, this complexity also makes it more challenging to target the construct of interest.***

In developing the assessment items, we drew on the work of Hill and her colleagues (2004, 2008) as well as science education researchers (Park & Oliver, 2008) that argue that PCK is highly contextualized in classroom practice. Consequently, we designed the assessment items embedded in vignettes about middle school science classrooms engaged in argumentation. In the advisors' reviews of the items, the use of the vignettes emerged as both a strength and a weakness in their overarching comments as well as their item specific feedback. For example, Advisor 10 wrote, "I think the greatest strength of the assessment items is that they are grounded in classroom science teaching scenarios… The use of scenarios also is a weakness in that the contextual features are not rich enough to make nuanced decisions about teaching practices. Several of the questions seemed like they had more than one appropriate response." Advisor 03 also wrote, "Well, the complexity of real teaching scenarios and multiple possibilities make

things muddy at times. A strength and a weakness. Not sure the actual constructs you are interested in are well isolated."

In providing item level feedback, for some questions the advisors felt there were multiple possible responses, because in an authentic classroom the context is important and there often is more than one "correct" option for teachers in terms of next steps. For example, Figure 4 includes Question 2 from the Mr. Strong vignette in which the students were addressing the question of whether or not humans could survive in settlements on Mars.

***Figure 4: Question 2 from Version 2 - Mr. Strong Vignette***

To get his students ready for the science seminar, Mr. Strong has them use the table to write arguments. Alicia and Thomas write the following arguments:

**Alicia:** I don't think humans can survive on Mars. The chart shows that Mars can get much colder than Earth and I saw a show on the Discovery Channel about the special clothes scientists have to wear when they do experiments in Antarctica because of the cold. It would be really awful to wear these clothes all the time just to go outside and it would cost a lot of money to get everyone these clothes.

**Thomas:** I think that settling on Mars would be great for humans. Days on Mars and Earth are almost the same length so we wouldn't have to change watches and clocks. Mars also has seasons like Earth so we'd have those too but they'd just be twice as long. Imagine how long summer break would be! No school for almost six months. Awesome.

2. After reading Alicia and Thomas's responses, Mr. Strong should:

a. Have students collect more numerical data about the planets under study
**b. Tell students to critique each other's claims about humans living on Mars**[1]
c. Ask students to analyze their current understanding of the scientific topic
d. Encourage students to better organize the evidence with a Venn diagram

[1] Correct answer choice is bolded.

The overall rating of this item was 2.5 which fell between aligns (2) and somewhat aligns (3); furthermore, one advisor selected the incorrect choice of d, resulting in this being one of the four items that we were required to revise (see Rules in Table 10). In their feedback, a number of advisors made the case that more than one of these choices would be appropriate next steps for Mr. Strong in terms of engaging students in critique (choice b), helping students with the science (choice c) and organizing their data (choice d). For example, Advisor 04 discussed both critique and the science content, "There appear to be several things that students need to do--relate the data to science more carefully being the main one--but this might be an outcome of critiquing each other's claims" while Advisor 07 discussed critique and the organization of the evidence, "I know you want me to say critique, but the individual arguments in the scenario are both badly organized from an evidentiary perspective." Consequently, in our revision of the item, we attempted to clarify the context in the opening of the vignette, the question stem, and the answer choices (see Methods S1 for Version 3). Overall, the advisors' feedback suggested that the focus on classroom vignettes was an appropriate direction; however, challenges also arise because of the inherent complexity of classroom instruction.

***Lesson #4: For the two dialogic argumentation conceptions, it was more challenging to develop high quality items and to distinguish between the two conceptions.***

In reviewing the advisors' feedback, we looked to see if there were any trends based on vignette and conception. Although all four vignettes received similar average ratings (1.88, 1.88, 1.925 and 1.725) between Aligns Very Well (1) and Aligns Well (2), we observed greater differences based on conceptions. The conception averages were: Conception 1: Evidence = 1.74, Conception 2: Reasoning = 1.54, Conception 3: Persuasion = 2.16 and Conception 4: Multiple Claims = 2.06. Both Conception 3 and 4, which focused on the dialogic aspects of argumentation, received weaker ratings averaging between Aligns Well (2) and Somewhat Aligns (3). The advisors' written feedback also reiterated the challenge of assessing these dialogic aspects. For example, Advisor 09 stated, "This dimension [Conception 3] and the fourth seem particularly hard to assess. A teacher could have multiple interpretations for the conversation that might not have to do with seeing persuasion as part of this." Furthermore, the advisors suggested that one of the reasons for the weaker ratings was the distinction between items designated as aligning with Conception 3 versus Conception 4 was unclear. For example, Advisor 05 wrote, "I think the conceptions are not mutually exclusive."

In reviewing Version 2 of the items, we agreed that it was not always clear why we labeled one item as Conception 3 and another item as Conception 4. For example, Mr. Strong Question 2 (see Figure 3) was labeled as Conception 4: Multiple Claims and the correct response was "Tell students to critique each other's claims about humans living on Mars." The next item for Mr. Strong, Question 3, was labeled as Conception 3: Persuasion and included the correct response, "The idea of a scientific argument is to convince everyone your claim is best." Although there are differences between these two items, in both cases multiple claims are being considered as well as students are engaging in questioning and critiquing. This makes sense since these are two aspects of dialogic argumentation that frequently occur simultaneously in the classroom, since the need to persuade an audience only arises if multiple claims are being considered. Consequently, we decided to revise our original four conceptions to include multiple claims and persuasion as two sub-goals of the overarching conception that "Students engage in dialogic interactions in which they try to convince an audience of the strongest among multiple claims." We still have the two conceptions explicitly labeled as sub-goals to remind ourselves to explicitly include both aspects in our continued revision of our assessment items. However, we are not necessarily trying to tease a part a distinction between these two goals. One assessment item can focus on both multiple claims and persuasion.

## Implications

PCK needs to shift from an abstract to a concrete construct to better support science teacher education and teacher professional learning (Berry et. al., 2008). With the recent focus on scientific practices in reform documents (NRC, 2012) and science standards (Achieve, Inc., 2013), we feel that it is important to explicitly articulate what the field means by PCK of scientific practices as well as how to both support and assess teacher learning of these important goals. In addition, PCK needs to be treated not as information, but considered in terms of how it manifests itself in action in a particular context (Settlage, 2013). Rather than viewing PCK as information teachers need to memorize and repeat back, we feel it is important to assess and support PCK in the context of k-12 instruction. Ultimately, our goal is to support teachers in developing PCK of scientific practices which they can use in action during their science

teaching. From our initial development process of an assessment for PCK of argumentation, we feel that this work has important implications for others designing or assessing teacher education programs for teachers at different stages of their career. Specifically, we offer recommendations in three areas: 1) Assessing and supporting a deep understanding of scientific practices, 2) Using PCK in classroom contexts and 3) Assessing and supporting argumentation as a dialogic process.

***Assessing and Supporting a Deep Understanding of Scientific Practices***

The development process for our assessment for PCK of argumentations highlights the challenge of developing a concrete measure of PCK. Answer choices that focused on argumentation, but were not the "best" answer for the question, were more effective distractors. In contrast, choices not as focused on the structural or dialogic aspects of argumentation, such as everyday experiences and vocabulary or procedural-related distractors, were not attractive to teachers, offering a limited measure of PCK. Furthermore, some distractors that used key argument terms like "evidence" and "persuasion" were still easily ruled out by teachers, because they focused on superficial aspects of the practice. One challenge of designing answer choices is ensuring the item assesses a deep understanding of the scientific practice. However, in revising the items to focus on more in-depth understandings, a new issue arose: for a few items there was not one clear correct answer selected by all of the advisors. Assessment items need to be carefully constructed focusing on characteristics of argumentation with the correct answer being specific enough for the context that experts in the field would clearly identify the choice.

In addition to assessing deep understandings of scientific practices, we also feel it is important to support such understandings in teacher education experiences. Teachers can feel that they are supporting scientific practices, but instead be focused on superficial features. For example, teachers can simplify argumentation such as turning the structural aspects into a formula or algorithm for their students (McNeill, 2009). Consequently, we suggest that the design of teacher education programs, such as in-service experiences, professional development and educative curriculum, should focus on introducing and supporting scientific practices within complex classroom contexts.

***Using PCK in Classroom Contexts***

Knowledge is not a collection of facts; rather, it is the activity an individual engages in that involves the person, tools and a context (Sawyer, 2006). Specifically for PCK, teachers develop and use knowledge within a given classroom context (Park & Oliver, 2008). Despite the importance of context, PCK often includes an implicit view of knowledge as "buckets" of information that teachers obtain, rather than a more activity oriented perspective focused on knowledge in use (Settlage, 2013). One of the challenges of developing an assessment for PCK of scientific practices, particularly one that can be given quickly and to a large number of teachers, is that it cannot occur in a real classroom. However, there is a need for the assessment to focus on teachers' application and use of knowledge (e.g. identifying a student's challenge with reasoning) rather than stating information (e.g. reasoning is often difficult for students). In our design, we used vignettes incorporating student writing and classroom transcripts to offer a classroom context for teachers to apply their PCK, but with the realization that it was an oversimplification of a real classroom. The advisors for our project identified the use of vignettes both as a strength, because there was a focus on classroom practice, but also a weakness. In terms of a weakness, two issues arose. First, the vignettes were not able to convey the richness of an actual classroom context. The second concern was in contrast to the first issue in that the

richness that was included at times resulted in more than one appropriate way for an individual to evaluate the situation.

We see the tension between these two concerns as being an issue not only for the design of assessments for PCK of scientific practices, but also for the development of teacher education experiences to support teacher learning. One potential avenue for supporting teacher learning is through the use of authentic records of classroom practice that utilize examples from k-12 classrooms, such as videos and student writing (Borko, 2004). The use of such images of practice has the same challenge in that teacher educators want to select them such that they highlight key aspects of classroom to support teacher learning; however, the images of practice should not be so simple that they lose the authentic complexity of classroom instruction.

***Argumentation as a Dialogic Process***

Specifically for argumentation, our development process suggested greater challenges with the dialogic aspects over structural characteristics of the practice. This finding aligns with previous research, which suggests that the dialogic aspects of argumentation may be particularly challenging for teachers (Alozie et al., 2010; McNeill & Knight, 2013; McNeill & Pimentel, 2010). Even when teachers are enacting curriculum with educative features supporting dialogic interactions, they can still rely heavily on traditional recitation discourse patterns that are primarily teacher directed (Alozie et al., 2010). In addition to the challenge of designing curriculum that support these interactions, our development work suggests that it is also challenging to develop teacher assessments that target PCK of this dialogic process. In our initial design, we attempted to target two separate conceptions related to the process – persuasion and multiple claims. However, the advisors' feedback suggested that these two conceptions may be too interrelated to assess as separate constructs. As such, our future development will focus on the dialogic characteristics from a more holistic perspective.

In addition, we feel that this finding has potential implications for teacher education as well. There is a tension and a messiness in trying to identify characteristics of argumentation in k-12 classroom to support teacher learning. On the one hand, labeling these characteristics can provide teachers with concrete elements to focus their attention; however, they can also distract teachers from considering argumentation as a more holistic scientific practice. Both types of lenses should be considered and applied when supporting teacher learning of argumentation.

**Acknowledgements**

This research was conducted as part of the Constructing and Critiquing Arguments in Middle School Science Classrooms: Supporting Teachers with Multimedia Educative Curriculum Materials project, supported in part by the National Science Foundation grant DRL-1119584. The design of the earth science curriculum was funded in part by a grant from the Bill & Melinda Gates Foundation. Any opinions expressed in this work are those of the authors and do not necessarily represent either those of the funding agencies, Boston College, Lawrence Hall of Science or the University of Berkeley. We would like to thank the teachers and advisors who provided valuable feedback on the items. Also, we would like to thank Lisa Marco-Bujosa for her assistance in the revision of the items.

**References**

Achieve, Inc. (2013). *Next Generation Science Standards.* Retrieved from - http://www.nextgenscience.org

Abell, S. K. (2007). Research on science teacher knowledge. In S. K. Abell & N. G. Lederman (Eds.). *Handbook of research on science education.* (pp. 3-28), Mahwah, NJ: Lawrence Erlbaum Associates.

Alozie, N. M., Moje, E. B. & Krajcik, J. S. (2010). An analysis of the supports and constrains for scientific discussion in high school project-based science. *Science Education, 94,* 395-427.

Baxter, J. A. & Lederman, N. G. (1999). Assessment and measurement of pedagogical content knowledge. In J. Gess-Newsome & N. G. Lederman (Eds.). *Examining pedagogical content knowledge.* (pp. 147-161). Dordrecht: Kluwer Academic Publishers.

Berland, L. (2011). Explaining variations in how classroom communities adapt the practice of scientific argumentation. *Journal of the Learning Sciences*, 20, 625-664.

Berland, L. K. & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education, 94*(5), 765-793.

Berland, L. K. & Reiser, B. J. (2009). Making sense of argumentation and explanation, *Science Education, 93*, 26-55.

Berland, L. K. & Reiser, B. J. (2011). Classroom communities' adaptations of the practice of scientific argumentation. *Science Education, 95*, 191-216.

Berry, A., Loughran, J. & van Driel, J. H. (2008). Revisiting the roots of pedagogical content knowledge. *International Journal of Science Education, 30*(10), 1271-1279.

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

Crippen, K. J. (2012). Argument as professional development: Impacting teacher knowledge and beliefs about science. *Journal of Science Teacher Education, 23*(8), 847-866.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372-400.

Hill, H. C. Schilling, S. G., Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal, 105*(1), 11-30.

Jiménez-Aleixandre, M. P. & Erduran, S. (2008). Argumentation in science education: An Overview. In S. Erduran & M. P. Jimenez-Aleixandre (Eds.). *Argumentation in science education: Perspectives from classroom-based research.* (pp. 3-28), Dordrecht: Springer.

Loper, S., McNeill, K. L., Peck, P., Price, J. & Barber, J. (2014, June). *Multimedia educative curriculum materials: Designing digital supports for learning to teach scientific argumentation.* Paper to be presented at the International Conference of the Learning Sciences, Boulder, CO.

McNeill, K. L. (2009). Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena. *Science Education, 93(*2), 233-268.

McNeill, K. L., Gonzalez-Howard, M. Katsh-Singer, R., Price, J. F. & Loper, S. (2013, April). *Teachers' beliefs and practices around argumentation during a curriculum enactment.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Puerto Rico.

McNeill, K. L. & Knight, A. M. (2013). Teachers' pedagogical content knowledge of scientific argumentation: The impact of professional development on k-12 teachers. *Science Education, 97*(6), 936-972.

McNeill, K. L., Lizotte, D. J, Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The Journal of the Learning Sciences, 15*(2), 153-191.

McNeill, K. L. & Pimentel, D. S. (2010). Scientific discourse in three urban classrooms: The role of the teacher in engaging high school students in argumentation. *Science Education, 94*(2), 203-229.

Moon, J., Passmore, C., Reiser, B.J., & Michaels, S. (in press). Beyond comparisons of online versus face-to-face PD: Commentary in response to Fishman et al., "Comparing the impact of online and face-to-face professional development in the context of curriculum implementation." *Journal of Teacher Education.*

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

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, 328, 463-466.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020.

Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C. & Richardson, K. (2013). Learning to argue: A study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching, 50*(3), 315-347.

Park, S. & Oliver, S. (2008). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education, 38*. 261-284.

Sadler, T. D. (2006). Promoting discourse and argumentation in science teacher education. *Journal of Science Teacher Education, 17*, 323-346.

Sampson, V., & Blanchard, M.R. (2012). Science teachers and scientific argumentation: Trends in view and practice. *Journal of Research in Science Teaching*, 49(9), 112-1148.

Sampson, V. & Clark, D. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*, 447-472.

Sampson, V. & Clark, D.B. (2011). A comparison of the collaborative scientific argumentation practices of two high and two low performing groups. *Research in Science Education, 41*, 63-97.

Sandoval, W. A. & Cam, A. (2011). Elementary children's judgment of the epistemic status of sources of justification. *Science Education, 95*, 383-408.

Sandoval, W. A. & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23-55.

Sawyer, R. K. (2006). Introduction: The new science of learning. In R. K. Sawyer (ed.), *The Cambridge handbook of the learning sciences.* (p. 1-16). New York, NY: Cambridge University Press.

Schneider, R. M. & Plasman, K. (2011). Science teacher learning progressions: A review of science teachers' pedagogical content knowledge development. *Review of Educational Research, 81*(4), 530-565.

Settlage, J. (2013). On acknowledging PCK's shortcomings. *Journal of Science Teacher Education*, 24, 1-12.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin Company.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

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

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, S. M. (2013). Professional development for science teachers. *Science, 340*(6130), 310-313.

Zembal-Saul, C. (2009). Learning to teach elementary school science as argument. *Science Education, 93*, 687-719.