The Development and Validation of a Learning Progression for Argumentation in Science

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A100692 *to Stanford University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education*

Abstract

Given the centrality of argumentation in the Next Generation Science Standards, there is an urgent need for an empirically validated learning progression of this core practice and the development of high quality assessment items. Here, we introduce a hypothesized three-tiered learning progression for scientific argumentation. The learning progression accounts for the intrinsic cognitive load associated with orchestrating arguments of increasingly complex structure. Our proposed learning progression for argumentation in science also makes an important distinction between construction and critique. We present validity evidence for this learning progression based on Item Response Theory, and discuss the development of items used to test this learning progression. By analyzing data from cognitive think-aloud interviews of students, written responses on pilot test administrations, and large-scale test administrations using a Rasch analysis, we discuss the refinement both of our items and our learning progression to improve construct validity and scoring reliability. Limitations to this research as well as implications for future work on assessment of scientific argumentation are discussed.

Keywords: argumentation, learning progression, assessment

Introduction

Argumentation is a central feature of science. As Crombie (1994, p. 3) states in the introduction to his three volume cognitive history of science, "The history of science in the European tradition is the history of vision and argument" – a view reflected in the central place argument occupies in the model of science offered in the *Framework for K-12 Science Education* (Fig 3.1, p. 45, National Research Council, 2012) which is the basis for the Next Generation Science Standards (Achieve, 2013). Argument and critique are, therefore, at the very center of science – connecting the "hands-on" work of scientific inquiry with the "minds-on" work of developing scientific ideas and theories.

Although student facility with scientific practices has been studied (Berland & McNeill, 2010; Grosslight, Unger, Jay, & Smith, 1991; Harrison & Treagust, 2000; Lehrer & Schauble, 2006; Schwarz et al., 2009), there is limited research around how to assess student ability with argumentation. In particular, few classroom assessments exist which specifically measure students' competence with argumentation, even though the study of argumentation has been a primary focus of much research in science education (Chang, Chang, & Tseng, 2010; Lee, Wu, & Tsai, 2009) and is a discourse practice which is seen as being central to learning and reasoning in science (Driver, Newton, & Osborne, 2000; Osborne, 2010). Therefore, if evidence-based argumentation is to become a common feature of science classroom, the field is faced with a challenge of a) defining progression in the competency of scientific argumentation, and b) providing items that assess student facility with this core practice.

A learning progressions approach

According to Corcoran, Mosher and Rogat (2009), “Learning progressions in science are empirically-grounded and testable hypotheses about how students’ understanding of, and ability to use, core scientific concepts and explanations and related scientific practices grow and become more sophisticated over time, with appropriate instruction” (p. 15). Such learning progressions outline potential cognitive paths that students might follow as they develop a more sophisticated understanding of a core scientific concept or practice. Not surprisingly, with the increasing emphasis on teaching scientific practices, in recent years there has been a resurgence of interest in the *progressions* or *trajectories* that students follow towards mastery of a specific domain or competency such as scientific argumentation.

Why the current resurgence of interest? First, there is the need to know how student ideas within a domain may become more sophisticated over time. This is not to argue that there is a single, linear route by which student conceptual understanding might develop—a common but flawed critique of learning progressions. Rather, learning progressions offer an informed understanding of the possible nature of student competency that have been empirically tested. Second, in recent years, more systematic and rigorous psychometric methods of testing such developmental constructs have emerged (Mislevy & Haertel, 2006; Wilson, 2013; Wilson & Sloane, 2000) using Item Response Theory (IRT). This approach offers the advantage of interpreting individual and discrete student measures in terms of a continuous spectrum of more general underlying traits spanning all persons, as well as providing more sophisticated estimates of the relative difficulty of tasks. Moreover, IRT data about student ability and item difficulty is not dependent on the sample at hand, which is a very attractive feature for researchers who seek to construct more general maps for how learning progresses in a particular domain or with a particular competency.

Why a learning progression for argumentation in science?

Despite extensive literature suggesting that argumentation supports student conceptual understanding (see Osborne, 2010 for a summary of the evidence), deliberative discourse of a dialectical nature where students engage in asking questions or elaborating and justifying their beliefs is rare in science classrooms (Banilower et al., 2013; Newton, Driver, & Osborne, 1999; Lemke, 1990). And, if reasoning and argumentation are to be emphasized in science classrooms, then such higher-order skills must be a feature of both the standards *and their assessment.* The latter, in particular, is important because the intentions of the curriculum are commonly read by teachers not from syllabi or curriculum documents, but from the items used to measure student achievement (Hannaway and Hamilton, 2008; Weiss et al., 2003). Transforming what we measure, however, is dependent on a common understanding of the construct of interest and its progression because, as Wiliam points out, "*assessments operationalize constructs*"[1] (2010, p. 259). However, little is known about how to readily *assess* reasoning and argumentation in a domain-specific manner, particularly in a quantitative way. Investigating scientific argumentation through the lens of a learning progression provides therefore a means of testing a

[1] Author's emphasis

developmental progression of this important practice and, in so doing, the compilation of a body of exemplar items that educators can use to assess scientific argumentation.

Our Model of Argumentation and its Development

Our view of scientific argumentation is that it is not simply an aptitude that can be assessed, but rather, a *competency* which draws on a mix of content knowledge, procedural knowledge, and epistemic knowledge (OECD, 2012). The construction of an argument requires the ability to remember appropriate information and to construct a justified relationship between this and the claim. As such it draws on two lower levels of the four types of knowledge in the revised taxonomy of Bloom's educational objectives (Anderson & Krathwohl, 2001) e.g., factual or conceptual, and may require the cognitive operation of constructing a dependent relationship between a claim and its supporting evidence or data. In contrast, critiquing an argument is somewhat more demanding requiring the cognitive operations of analysis to identify the salient elements of an argument i.e., claim, warrant, data, followed by an evaluation of the truth status of these elements or their validity while drawing on factual or conceptual knowledge, and then creating or synthesizing a counter-argument which is relevant to the argument that has been advanced. Undertaking such a process is aided by a metacognitive knowledge (the highest level in the revised taxonomy) of the nature of an argument, and an ability to distinguish its component elements. Thus, in developing our model of a progression in argument, we hypothesized that the process of critique would be more cognitively demanding than the process of constructing an argument. Such a position is supported by the evidence that the reasoning of individuals suffers from confirmation bias in that they are much more readily able to justify the claims they wish to advance than to engage in more the cognitively demanding process of constructing counter-arguments or potential refutations (Koslowski, Marasia, Chelenza, & Dublin, 2008; Mercier & Sperber, 2011;Wason & Johnson-Laird, 1972).

Argument is also different from explanation (Osborne & Patterson, 2011). The latter seeks to identify dependent relationships answering, whenever possible, the causal question of why a particular phenomenon occurs. Moreover, the goal of explanation is understanding. In contrast, the goal of argument is persuasion – and scientific argument is a complex form of reasoning requiring domain-specific knowledge to construct and critique claims and their relation to any supporting evidence (VonAufschnaiter, Erduran, Osborne, & Simon, 2008; Osborne, Erduran & Simon, 2004) to persuade other members of the community of their validity. Thus the nature of argumentation is such that it requires that the assessment of student competency within a specific domain. In our case, we have chosen to ground this study in the domain of the structure of matter, given its centrality to all sciences. Situated within this domain, our primary research question is:

1. How do middle school students demonstrate competence with increasingly complex argumentation in the context of the structure of matter?

Additionally, a second and subsidiary question of interest concerns the potential interaction between the complexity of an argument and the domain in which that argument is situated. We offer a hypothesized learning progression for science argumentation, but we test it specifically in the structure of matter context, which leads us to ask:

2. To what extent is student competence with increasingly complex argumentation generalizable?

The learning progression used in this research was developed using what was described above as a “cognition and instruction” or “top-down” approach. That is, our hypothesized progression began with research-driven hypotheses about how students might develop the ability to argue in increasingly sophisticated ways. Then, assessment items were developed to specifically probe levels of our hypothesized progression. While any learning progression cannot be a one-size-fits-all answer to how student argumentation increases in complexity (Shavelson and Kurpuis, 2012), our hypothesized learning progression can provide a framework for teachers to gauge the level at which their students engage in argument from evidence, as well as estimate the relative complexity of the argumentative tasks and assessments they may require of their students. Furthermore, our progression has the potential to inform policy makers and curriculum developers on how to better align science content standards and assessments with what students can realistically achieve.

Previous work on developing learning progressions for scientific argumentation

Berland and McNeill (2010) proposed a learning progression for scientific argumentation consisting of three dimensions – *instructional context*, *argument product*, and *argument process* characterizing the ways in which students’ arguments vary in complexity and sophistication across grade levels and instructional contexts. Their findings indicate that, with guidance, students as young as 5th grade can engage in meaningful argumentation. Furthermore, this work provides initial evidence that argumentation ability develops over time, as the structure of argumentation was more complex in 12th grade students' work. However, their evidence for their progression was drawn from qualitative classroom observations and they did not systematically assess student competency with scientific argumentation as defined by the levels of their learning progression.

Songer, Gotwals, and colleagues have built an extensive body of work investigating the development of complex thinking in biology (Songer, Kelcey & Gotwals, 2009; Gotwals & Songer, 2010; Gotwals & Songer, 2013). In more recent publications, they have divided the "complex thinking" construct into two separate progress variables: core disciplinary ideas (with strands for classification, ecology and biodiversity) and constructing an evidence-based scientific explanation. Their progress variable is called "constructing evidence-based explanations," although the learning progression is essentially a measure of a student's ability to construct an argument, focusing as it does on a students' ability to advance a claim, support it with appropriate evidence, and use reasoning to link the claim and the evidence, with and without the use of written scaffolds.

Finally, Lee et al. (2014) have identified "level of uncertainty" as a potentially important facet of the practice of argumentation, and proposed a learning progression for incorporating uncertainty into scientific argumentation. Their proposed learning progression consisted of "making a claim" at the base, followed by "providing justification," "identifying uncertainty," and "providing justification for uncertainty." They found that the recognition of uncertainty was not aligned with their other three elements, and hence eliminated "identifying uncertainty" from further analysis. Their subsequent analysis provided support for the structure of their learning progression – average item difficulties increasing in the order of claim, justification, and the rationale for uncertainty. However, Lee et al. found that rather than justifying the uncertainty of the argument, many students were uncertain because they did not understand the scientific content. Hence Lee et al. were not able to distinguish "uncertainty" in the logical structure of arguments from "uncertainty" about the domain specific content in which the arguments were situated. In contrast, we argue beneath that differentiating between the act of "construction" and the act of "critique" provides a more generalizable approach to describing the development of argumentation than does offering a justification for uncertainty.

Argumentation as a process of construction and critique

Argumentation requires the ability to engage in both construction and critique - in particular, because the construction of knowledge is a dialectic between construction *and* critique (Ford, 2008). Construction attempts to use argument to defend and support an explanatory hypothesis, for instance the classification of a species or a sample, or to defend a particular interpretation of data or experimental design. Critique, in contrast, is an attempt to establish why an argument is flawed or incorrect. Claims do not exist in isolation—instead, they exist in competition with other ideas (Howson and Urbach, 2006; Allchin, 2012; Longino, 1990). Hence critique is not merely a peripheral activity for scientists, but essential to the establishment of a deeper and improved understanding of the material world. Mercier and Sperber (2011) point to multiple empirical studies suggesting that individuals suffer from confirmation bias and fail to anticipate the counter-arguments. Critique of others' ideas promotes epistemic vigilance and improves the quality of reasoning. Hence, given its central epistemic role, it is surprising that so little attention has been paid to mapping students' competence with critique. And since argumentation is a "verbal, social and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or *refuting*[2] the proposition expressed in the standpoint" (Van Eemeren & Grootendorst, 2004, p. 1), a complete performance of the practice should require students to engage in critique. Hence, our proposed learning progression for argumentation differs from previous work in that we operationalize argumentation as a combination of *both* construction *and* critique.

Our Proposed Learning Progression for Argumentation

Towards a working definition of argumentation

Competencies, such as argumentation, draw on a mix of content knowledge, procedural knowledge, epistemic knowledge and skill (OECD, 2012) and can be seen as "context-specific cognitive dispositions that are acquired and needed to successfully cope with certain situations or tasks in specific domains" (Koeppen, Hartig, Klieme, and Leutner, 2008, p. 62). Building on this conception of competency, we view *scientific argumentation* as a complex process of reasoning utilized in situations that require scientific content knowledge to construct and/or critique proposed links between claims and evidence.

[2] Emphasis added.

For our hypothesized progression, we have drawn on Toulmin's (1958) model for the structure of *practical* or informal arguments. Toulmin's model (1958) begins with a *claim* as a "conclusion whose merits we are seeking to establish" (p. 90) which is supported by relevant *data or evidence*. The relation between the evidence and the claim is provided by a *warrant* that forms the substance of the justification for the claim. Warrants may often be dependent on implicit assumptions that Toulmin referred to as *backing*. Claims may also be circumscribed by the use of *qualifiers* that define the limits of validity. Because of its relative simplicity, Toulmin's practical model has formed the basis of many schemas used in research analyzing student discourse (Cavagnetto, 2010; Erduran, Simon, & Osborne, 2004; Zohar & Nemet, 2002).

While Toulmin's (1958) model of practical argument plays a central role in our learning progression for argumentation, it is not sufficient. In addition to justification, the process of *critique* is essential to identifying flaws in arguments. As discussed previously, such competency is reliant on both a knowledge of content, a tacit meta-knowledge of the features of an argument, the ability to distinguish its component elements, and to construct a rebuttal. Commonly, this may require the cognitive performance of comparing and contrasting the relative merits of two arguments or constructing an argument for why some evidence has higher epistemic validity than other evidence (Koslowski et al., 2008; Zimmerman, 2007). In addition, this mean that there is a domain-specific element to the competency required to engage in scientific argumentation and reasoning which is context and task dependent. (Tricot and Sweller, 2014, Fischer et al, 2014).

In summary, we conceptualize argumentation as a competency whose goal is the resolution of one or more claims (Walton, 1990) where resolution is only possible by engaging in a process of critique, which is essential for identifying flaws in any arguments (Ford, 2008; Henderson, Osborne, MacPherson, & Wild, 2015). Therefore, taken as a whole, a competency for scientific argumentation demands a complex orchestration of construction and critique of claims, warrants, and evidence in situations that require scientific knowledge to resolve. Hence, argumentation is both a novel and demanding task for many students that can – depending on the complexity of the argument – make substantial cognitive demands.

Accounting for cognitive load

Like any cognitive task, argumentation makes demands on working memory. Cognitive Load Theory (CLT) (Chandler & Sweller, 1991; Paas, Renkl, & Sweller, 2003; Sweller, 1994) assumes that human working memory capacity is limited, and provides a framework for the design of instructional materials and assessments Pollock, Chandler, and Sweller (2002) summarize two basic sources of cognitive load. The first is *intrinsic cognitive load,* which "is determined by the extent to which various elements interact… An element is the information that can be processed by a particular learner as a single unit of working memory" (p. 62). In contrast, the second is extraneous cognitive load, which "is generated by the manner in which information is presented to learners and is under the control of instructional designers" (p. 62).

For the purposes of this study, CLT offers a means by which the competency of scientific argumentation can be operationalized as a single progress variable – the intrinsic cognitive load on student working memory imposed by differing levels of coordination required between the core elements of arguments. Specifically, the process of constructing or critiquing arguments can be viewed as an orchestration of various combinations of claims, warrants and data. Hence, it follows that for tasks that increase the number of element that must be coordinated, the intrinsic cognitive load on the working memory increases, thereby making it more difficult to demonstrate argumentative competency. Thus cognitive demand provides a foundation for a single process variable by which scientific argumentation can be systematically assessed – in contrast to previous depictions in the literature where scientific argumentation has been conceptualized as "messy" when defined by multiple progress variables (Gotwals, Songer, & Bullard, 2012).

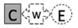
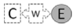
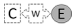
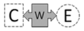
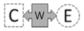
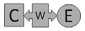
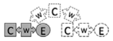
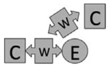
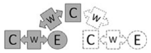
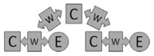
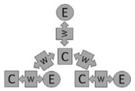
The notion of cognitive load also assists in constructing a realistic top anchor for our learning progression, by limiting assessment items to those which are within the limits of human working memory. Finally, CLT has also guided our creation of assessment items to test each level of our argumentation learning progression by focusing our attention on the intrinsic cognitive demands of Toulmin's elements of argument while seeking to minimize the *extraneous cognitive load* of how our assessments are formatted. For example, we worked with teacher co-developers to flag and remove potentially confusing words, figures, or images from the assessments we have designed to test our proposed learning progression.

Our learning progression

Our hypothesized learning progression for argumentation is shown in Table 1 and consists of three broad levels of argumentation differentiated by intrinsic cognitive load, where each level is seen as requiring more connections to be made between claims and pieces of evidence. Early levels are prefixed with the number zero to denote that assessment items probing these stages do not ask for explicit connections between claim and evidence to be made. These connections – warrants under the Toulmin model – are not specifically asked for, and hence it is possible to demonstrate competency with identification/critique of an isolated claim, warrant, or evidence without making a logical connection between them. Hence the zero prefix for these levels denotes *zero degrees of coordination*.

Items that require the construction of relationships between claims and evidence (i.e., warrants) mark the transition from Level 0 to Level 1 on our learning progression. These intermediate levels are prefixed with the number one to denote *one degree of coordination* – i.e., students need to make one explicit logical connection between claim and evidence by way of a warrant. This level builds on Level 0 of the learning progression, as it requires understanding of not only what constitutes a claim or a piece of evidence, but also how to construct or critique a relationship between claim and evidence. The most advanced levels of our learning progression are prefixed with the number two to denote *two or more degrees of coordination*. Level 2 items require students to explicate or compare two or more warrants. An elaborated description of this learning progression, which cross-references Table 1 to provide concrete examples of how various progress levels are operationalized with assessment items, is included in the Supplementary Materials.

A proposed learning progression for scientific argumentation

Level	Constructing	Critiquing	Description	Representation of Elements
0			No evidence of facility with argumentation.	
0a	Constructing a claim		Student states a relevant claim.	
0b		Identifying a claim	Student identifies another person's claim.	
0c	Providing evidence		Student supports a claim with a piece of evidence.	
0d		Identifying evidence	Student identifies another person's piece of evidence	
1a	Constructing a warrant		Student constructs an explicit warrant that links their claim to evidence.	
1b		Identifying a warrant	Student identifies the warrant provided by another person.	
1c	Constructing a complete argument		Student makes a claim, selects evidence that supports that claim, and constructs a synthesis between the claim and the warrant.	
1d	Providing an alternative counter argument		Student offers a counterargument as a way of rebutting another person's claim.	
2a	Providing a counter-critique		Student critiques another's argument. Fully explicates the claim that the argument is flawed and *justification* for why that argument is flawed.	
2b	Constructing a one-sided comparative argument		Student makes an evaluative judgment about the merits of two competing arguments and makes an explicit argument for the value of *one* argument. No warrant for why the other argument is weaker.	
2c	Providing a two-sided comparative argument		Student makes an evaluative judgement about two competing arguments and makes an explicit argument for why one argument is stronger and why the other is weaker.	
2d	Constructing a counter claim with justification		This progress level marks the top anchor of our map. Student compares and contrasts two competing arguments, and constructs a new argument in which they justify why it is superior to previous arguments.	

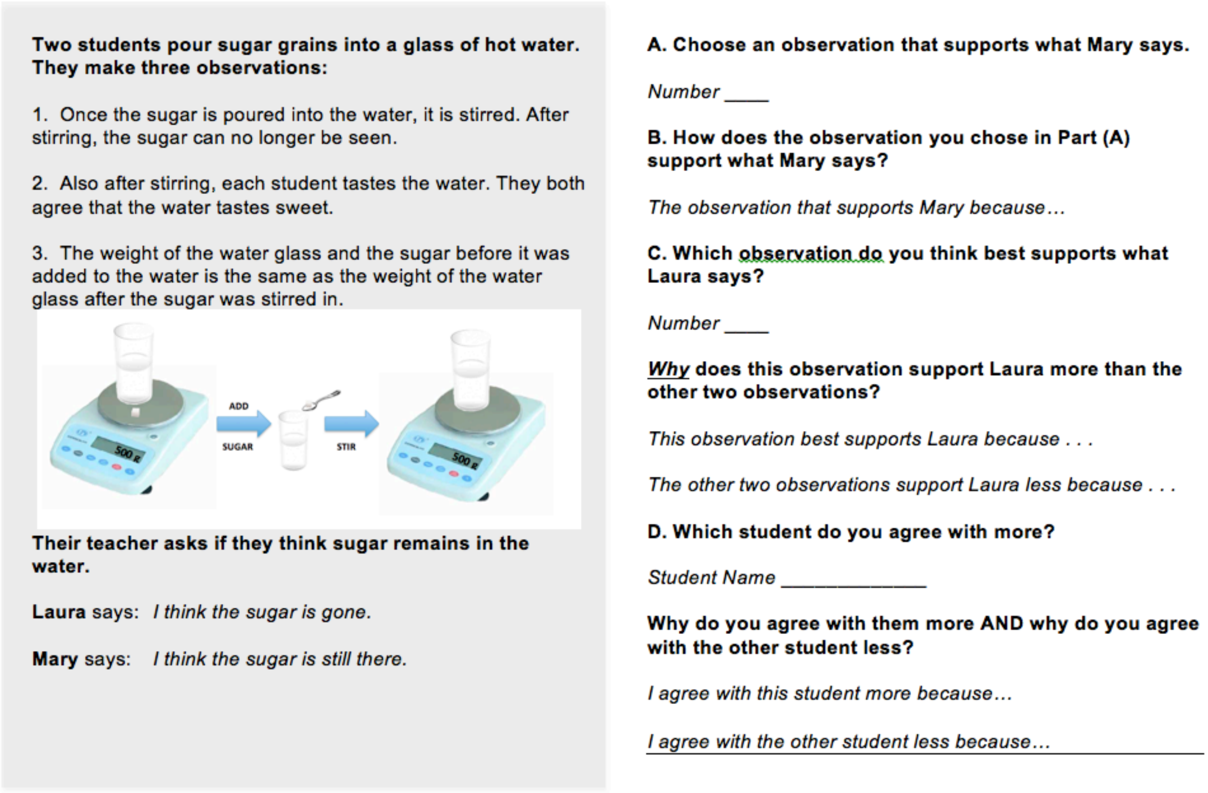
Note: Symbolically, the greater intrinsic cognitive load commensurate with advanced levels of our learning progression is represented by argument diagrams of increasing complexity. When both a construction (e.g., Level 0c) and critique (e.g., Level 0d) tasks have the same degree of complexity, the critique task is considered more advanced because it is more difficult for middle school students to identify somebody else's warrant than it is for them to construct their own.

We

hypothesized that items testing higher levels will pose greater difficulty for students to demonstrate competence than for items testing lower levels. As our proposed learning progression operationalizes argument as consisting of both construction and critique, our progression distinguishes arguments that are originally constructed by a student from critiques of arguments constructed by someone else – a distinction represented by the columns for Constructing and Critiquing. For any given task, we hypothesized that if a student coordinates the elements of their own argument, this will be less difficult than the more abstract task of critiquing another's argument for the reasons given above. This was confirmed by our pilot studies which suggested that it was easier for students of middle school age to advance their own thinking than to offer and/or critique others' arguments.

To aid in the visualization of the increasing complexity of argument elements that must be coordinated as progress levels advance, our learning progression contains diagrams that depict the structure of arguments at various progress levels. In these diagrams claims are denoted by squares, evidence denoted by circles, and the warrants used to coordinate claims and evidence are denoted by a box with connecting arrows. Furthermore, these diagrams differentiate what students must specifically *explicate* on items probing each level of our learning progression as opposed to those elements which may be implicitly demanded. Shaded elements reflect what students are expressly asked for in an item testing a particular level of the learning progression. For example, when an item prompts a student to explicate an argument in full (Level 1c of our learning progression), then all elements are shaded in the diagram. When some diagram elements are not shaded, the item does not ask for all aspects of that argument to be explicated, for instance, when an item asks a student to identify a claim in an argument that is not their own (Level 0b of our learning progression). Here the student only needs to explicate the claim (shaded in the argument diagram for Level 0b) to receive full credit, but to do so they may still consider what aspects of the argument are *not* the claim (i.e., the evidence and the warrant are unshaded in the argument diagram for Level 0b). This shading scheme in our argument diagrams is meant to reflect our decision to assess only what is *explicitly* asked of students taking the items we created.

In addition to the learning progression, we provide an example of a bundle of assessment items used to probe different levels of our learning progression (Figure 1). In the scenario depicted in Figure 1, three pieces of evidence are presented about what happens when sugar is stirred into a glass of water. In addition to this evidence, two differing claims are posed by fictitious characters. This requires students to draw on their scientific content knowledge to construct and/or critique proposed links between claims and evidence.



Two students pour sugar grains into a glass of hot water. They make three observations:

1. Once the sugar is poured into the water, it is stirred. After stirring, the sugar can no longer be seen.

2. Also after stirring, each student tastes the water. They both agree that the water tastes sweet.

3. The weight of the water glass and the sugar before it was added to the water is the same as the weight of the water glass after the sugar was stirred in.

Their teacher asks if they think sugar remains in the water.

Laura says: *I think the sugar is gone.*

Mary says: *I think the sugar is still there.*

A. Choose an observation that supports what Mary says.

Number ____

B. How does the observation you chose in Part (A) support what Mary says?

The observation that supports Mary because…

C. Which observation do you think best supports what Laura says?

Number ____

Why **does this observation support Laura more than the other two observations?**

This observation best supports Laura because . . .

The other two observations support Laura less because . . .

D. Which student do you agree with more?

Student Name ____________

Why do you agree with them more AND why do you agree with the other student less?

I agree with this student more because…

I agree with the other student less because…

Method

Sample

The development of our argumentation items and learning progression was conducted with middle school students from a large school district in the San Francisco Bay Area. Middle school students were chosen as the topic of structure of matter, the domain focus, is addressed in some detail in grades 6-8. Second, research would suggest that it is at this age that students are more likely to be developing an evaluative disposition towards knowledge claims (Kuhn, 1999) and are less reluctant to engage in argument (Kuhn, Wang, & Li, 2011). At the time of the research the district had 55,091 students, of which 10,131 were middle school students coming from a wide diversity of race and cultures. Over the course of the entire four-year project, approximately 2,000 students were tested or interviewed as part of this project and drawn from the full range of schools and abilities which this district serves. This paper reports on data collected during the 2012-2013 school year from 803 students. Sixteen (8 males, 8 females) additional students participated in think aloud interviews with a set of argumentation test items. A follow-up investigation was conducted in 2014, which included both 8th graders and 10th graders. The data for participants in this study are summarized in Table 2.

Year	Activity	Sample
2013	Written assessment items	**803** eighth-grade students
	Think aloud Interviews	**16** eighth-grade students (8 female; 8 male)
2014	Written assessment items	**119** eighth-grade students; 159 tenth-grade students
	Think aloud interviews	**15** eighth-grade students

Table 2: Summary of participants in each phase of the research. Students who took the written assessment items completed one of four different test forms containing scientific items, general items, or both types of item. Participants who completed interviews were selected by their classroom teachers.

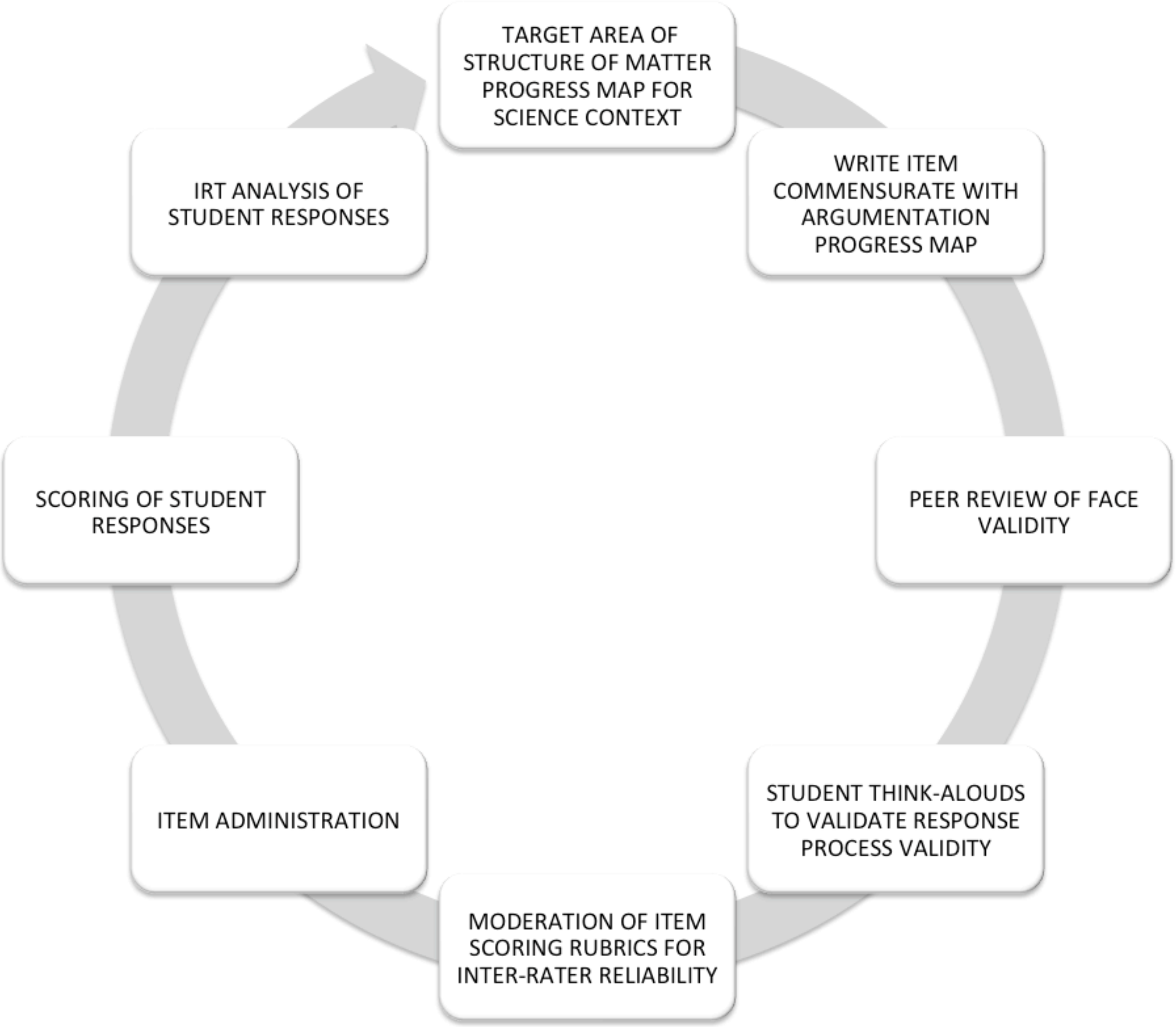
Assessment development

The approach to testing our learning progression was based on the *BEAR Assessment System* (Wilson & Sloane, 2000), operationalized through four primary building blocks: the construct map (of argumentation), items design, outcome spaces, and the measurement model (Wilson, 2013). The *construct map* represents the latent construct being probed. In the BEAR system, the developmental perspective emphasizes how students progress from lesser to greater expertise in the domain of interest. *Items design* refers to the systematic design of tasks to elicit specific kinds of evidence about student knowledge, as described in one or more construct maps. *Outcome spaces* define the qualitatively different levels of responses (of the construct map) relative to a particular prompt or stimulus; essentially this is where a value is placed on student work. Finally, the *measurement model* defines how inferences about student understandings are drawn from the evaluated (scored) work. In the BEAR system, we use a Rasch-based model known as the partial credit model (Masters, 1982) and its generalization the multidimensional random coefficients multinomial logic model (MRCMLM) (Wilson & Wang, 1995). The models provide convenient and rich ways to model both person proficiency and item difficulty on the same scale. In addition, items from different types of assessments can be scaled together so that student gains can be evaluated in a straightforward way without requiring students to take the same pre- and post-test.

All assessments for this project went through a rigorous development and evaluation process that followed several iterations of the four steps mentioned above. Using Rasch models allowed for an investigation of the measurement properties of the test by estimating an individual's ability, his/her fit statistics and the item difficulties and the item fit statistics. For this research, individual ability was interpreted as by individual's proficiency in response to the items on a continuum as represented on the learning progression. "Item difficulty" was interpreted as the degree to which typical individuals generate or choose the correct answer to an item. Rasch analysis is useful because it translates raw information (item scores and person estimates) into a logit scale, placing item locations and individual ability estimates on one scale using equal intervals of relative difference. Conducting this transformation of the data places items, subject to fit testing, on a new scale free from the particulars of persons or items (Masters, 1982). This analysis was done using ConQuest software (Wu, Adams, & Wilson, 2012). The software provides visual maps and results that can be interpreted using the learning progression as a guide. Comparing the response patterns of the students' responses to the items with the predictions from the hypothesized model can then illuminate discrepancies in the "theory" or learning progression, or alternatively, in the item design and implementation.

To build coherent and evidence-based learning progressions, it is necessary to ensure the quality of the assessments. Such evidence is best sought from a close examination of the individual respondent and his/her individual responses (Wilson, 2013). Messick (1987) writes that it has long been presumed that the score a person attains on a test is determined by relevant responses to the specific content and, in addition, it is usually taken for granted that this score reflects the respondent's knowledge of the domain tested by their achievement on tests (for example). However, he suggests that, "evidence in support of these presumptions is critical in establishing the meaning or construct validity of the scores" (Messick, 1991, p. 161). Therefore, verbal report protocols (Ericson and Simon, 1993) combined with an exit interview were administered to a selected sample of 10-20 students per question. A verbal report protocol was used to audio-record students who were asked to complete the test while concurrently "thinking out loud", occasionally with minimal prompting (e.g. "please keep talking", "I'm listening".) Data from these protocols was used to refine and improve the items.

Significant steps were taken to ensure inter-rater consistency during the scoring of the assessments. First, the scoring guide was exemplified with typical student responses that were taken from pilot administrations of the assessments. A group of trained raters "moderated" each scoring guide to ensure that there was high agreement about what constituted evidence of each score. Items were then rated using the scoring guide. Ratings and assessments were then shared. Disagreements were resolved through discussion and the scoring guides amended appropriately. The complete cycle of development described above is summarized in Figure 2. We repeated this cycle several times to reach the final item set. Final inter-rater reliability was 0.74 based on overlapping scoring of 20% of the items by an independent rater. This reliability is considered reasonable given the innovative nature of the items which produced a wide variety of student responses and therefore some variation in interpretation among the coders.

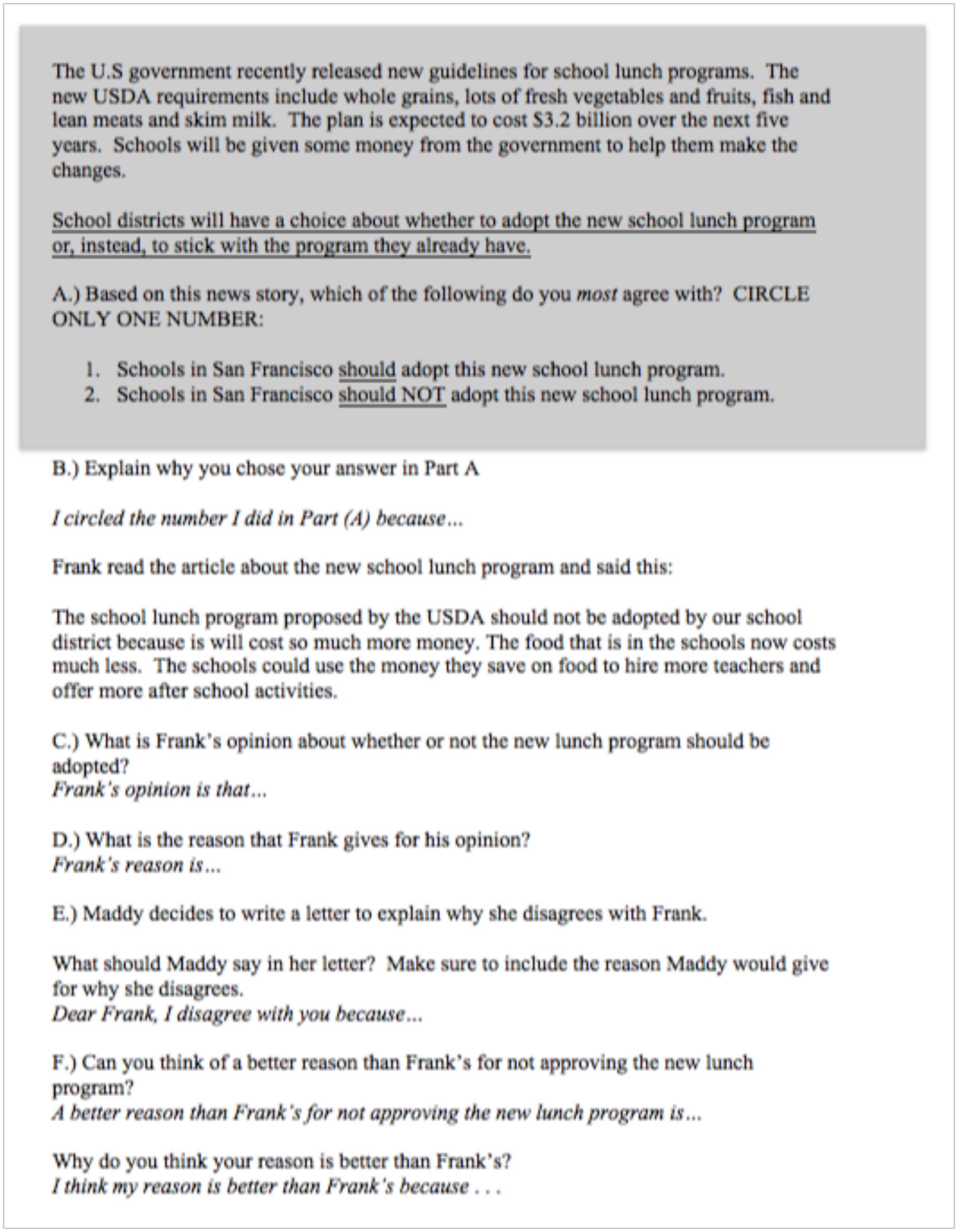

TARGET AREA OF STRUCTURE OF MATTER PROGRESS MAP FOR SCIENCE CONTEXT
WRITE ITEM COMMENSURATE WITH ARGUMENTATION PROGRESS MAP
PEER REVIEW OF FACE VALIDITY
STUDENT THINK-ALOUDS TO VALIDATE RESPONSE PROCESS VALIDITY
MODERATION OF ITEM SCORING RUBRICS FOR INTER-RATER RELIABILITY
ITEM ADMINISTRATION
SCORING OF STUDENT RESPONSES
IRT ANALYSIS OF STUDENT RESPONSES

To summarize, in each cycle of development we: (1) picked a content progress variable in which to situate a given argumentation item; (2), wrote an initial draft of the item; (3) conducted peer review within our research group (composed of argumentation and content experts) and with teacher co-developers of each item; (4), conducted think aloud interviews with eighth grade students to establish construct validity and identify other item flaws; (5) moderated detailed scoring guides for items based on pilot test administrations; (6) administered final items to eighth grade students; (7) scored student responses based on the moderated item rubrics; and (8) used Rasch analysis to analyze findings from large-scale administrations.

Description of the final set of assessment items

A set of argumentation items contextualized in the physical behavior of matter was designed in paper and pencil format. Similar to other performance assessments attempting to measure higher order thinking skills, the argumentation assessment in this study consisted of different tasks, for which a setting was established and multiple questions related to the setting were presented subsequently. Using the Sugar in Water item (Figure 1) as an example, groups of items were situated in a common context consisting of (1) a scientific question (e.g. "What happens when you stir sugar into a glass of hot water?) (2) hypothetical evidence (e.g. "When the water is stirred, the sugar can no longer be seen") and (3) different arguments addressing the question. This made various types of questions possible, including tasks where students critique the fictitious arguments and/or construct their own arguments that they believe to be superior. Rosenbaum (1988) introduced the term *item bundle* to denote a subset of items that share a common stimulus. In the remainder of the text, items sharing the same scenario will be referred to as an "item bundle."

Over the course of three years, 21 item bundles assessing scientific argumentation were developed, tested, and analyzed psychometrically. This complete set consisted of a total of 93 individual items. In the third year of the project a set of three "general argumentation" item bundles were developed. Our intent for the contexts of the "general" items was that they would be familiar enough to 8th grade students so as to not require any domain-specific knowledge to construct or critique the arguments. An example of a “general” assessment item bundle is “School Lunch,” which is shown in Figure 3: Other examples can be found at http://scientificargumenation.stanford.edu



The U.S government recently released new guidelines for school lunch programs. The new USDA requirements include whole grains, lots of fresh vegetables and fruits, fish and lean meats and skim milk. The plan is expected to cost $3.2 billion over the next five years. Schools will be given some money from the government to help them make the changes.

School districts will have a choice about whether to adopt the new school lunch program or, instead, to stick with the program they already have.

A.) Based on this news story, which of the following do you *most* agree with? CIRCLE ONLY ONE NUMBER:

1. Schools in San Francisco should adopt this new school lunch program.
2. Schools in San Francisco should NOT adopt this new school lunch program.

B.) Explain why you chose your answer in Part A

I circled the number I did in Part (A) because…

Frank read the article about the new school lunch program and said this:

The school lunch program proposed by the USDA should not be adopted by our school district because is will cost so much more money. The food that is in the schools now costs much less. The schools could use the money they save on food to hire more teachers and offer more after school activities.

C.) What is Frank's opinion about whether or not the new lunch program should be adopted?
Frank's opinion is that…

D.) What is the reason that Frank gives for his opinion?
Frank's reason is…

E.) Maddy decides to write a letter to explain why she disagrees with Frank.

What should Maddy say in her letter? Make sure to include the reason Maddy would give for why she disagrees.
Dear Frank, I disagree with you because…

F.) Can you think of a better reason than Frank's for not approving the new lunch program?
A better reason than Frank's for not approving the new lunch program is…

Why do you think your reason is better than Frank's?
I think my reason is better than Frank's because . . .

The Rasch analysis presented in this paper for the purpose of supporting our hypothesized learning progression is based on the complete 2012-2013 argumentation assessment data set, which consisted of 4 scientific argumentation item bundles (20 individual items) and 3 general argumentation item bundles (15 individual items). These items are from the third year of the project and thus incorporate many of the suggestions from earlier cycles of development. The full set of items can be found in Supplementary Materials.

Findings

Item statistics

Basic psychometric properties of the scientific argumentation and general argumentation items were calculated and are reported in full in the Supplementary Materials. Item discrimination indicates the extent to which success on an item corresponds to success on the whole test, and is thus a measure of how well an item differentiates between students of higher and lower ability. The classical item discrimination index can have a value between 0 and 1, with values closer to 1 indicating higher discrimination. Generally, values between 0.3 and 0.7 are acceptable on tests of academic achievement (Osterlind, 1998). Discrimination index values for this set of items ranged from .34 to .66, which is within the acceptable range. We used weighted mean squared residual (MNSQ) as a measure of Rasch item fit, a parameter which is based on the difference between what is observed and what is predicted by the model. An acceptable value for the weighted MNSQ is between 0.75 and 1.33 (Wilson, 2005). All items fell well within an acceptable range for discrimination and weighted mean square. Therefore, we proceeded with the rest of the unidimensional and multidimensional Rasch analyses, and used the results from these analyses to make conclusions about the relative difficulty of items and, thus, the development of argumentation competence.

Dimensionality of an assessment containing scientific and general items

We next investigated whether the items situated in a scientific context (i.e., scientific argumentation) and items situated in a general (familiar) context (i.e., general argumentation), analyzed together, fit better to a unidimensional model or a multidimensional between-item model. Thus, following the separate unidimensional analyses of both the scientific argumentation and the general argumentation items, we conducted a multidimensional analysis of the scientific and general argumentation items. 490 of the 803 total students in the 2012-2013 sample took *both* scientific argumentation and general argumentation items. The multidimensional model fit for those students fits significantly better than the unidimensional model (*df* = 2, chi-square = 67.44, *p* < .001).

Reliability	**Scientific Dimension**	**General Dimension**
MLE person separation	0.84	0.71
WLE person separation	0.84	0.69
EAP/PV	0.87	0.84

Table 3: Reliability coefficients for maximum likelihood estimation (MLS) person separation, weighted likelihood estimation person separation, and expected a posteriori estimation based upon plausible value (EAP/PV)

This finding indicates that the scientific argumentation items and the general argumentation items tested different concepts. The correlation between the two dimensions was strong and positive (0.86) and the reliabilities of the two dimensions found to be 0.87 and 0.84, respectively (Table 3). This finding suggests that the scientific argumentation items required a combination of both content knowledge in the structure of matter domain as well as competency with the construction and critique of evidentiary arguments. In contrast, the general argumentation items were situated in everyday contexts more likely to be familiar to the students participating in this study, and hence success on these general argumentation items did not require domain-specific knowledge as with the scientific argumentation items.

Comparing the difficulties of argumentation items across scientific and general contexts

The first research question asked, “How do middle school students’ demonstrate competence with increasingly complex argumentation in the context of the structure of matter?” The Wright map from the multidimensional between-item model (Figure 4) provides evidence that answers this question. The Delta dimensional alignment technique (Schwartz & Ayers, 2011) was applied to calibrate the scientific argumentation and general argumentation dimensions onto a common metric. As such, the estimates of the ability and item parameters of these two dimensions can be compared. The left panel of the Wright map shows the ability distribution and item thresholds of the scientific argumentation dimension, while the right panel shows the ability distribution and item thresholds of the general argumentation dimension. The Wright map shows how items and students were distributed on the argumentation scale ranging from -3 to +3 in logit values. The distribution of students' ability estimates is shown by the X's to the left. The higher the students' performance on the assessment, the more able the student is on the scientific argumentation construct. Item threshold values are indicated by shaded boxes. An item threshold is defined as the value at which students with the matching ability would have a 50% chance of receiving a score of X "or above" as compared to receive a score "below X". In this figure, only the threshold for for students receiving full credit is displayed.

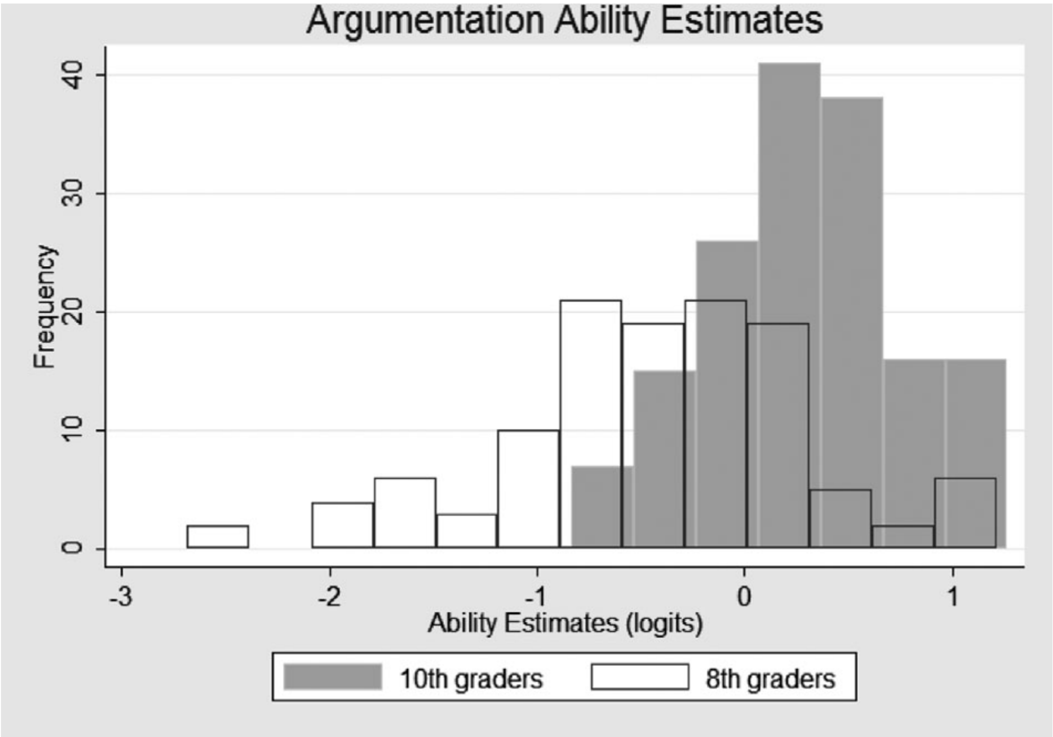
	Scientific Argumentation				General Argumentation				
Logit	Ability	Level 0	Level 1	Level 2	Ability	Level 0	Level 1	Level 2	Legend
3									**B**:
									"Bubbles
									in Water"
									O:
									"Onions"
2								V_de	
									G:
									"Facts
				B_6					About
					x			F_fg	Gases"
	x			O_5	x				
	x			O_6	x			L_e	**S**:
	x			G_cd G_e	xx			F_hi V_c	"Sugar
	xx				xxx			L_f	Dissolving
	xx			S_d	xxx				in Water"
1	xxxx			S_c	xxx				
	xxxx		O_7		xxx				**L**:
									"School
	xxxxx				xxxxxx				Lunch"
	xxxxxx		O_5	B_5	xxxxx				
	xxxxxxxxxx				xxxxxx				**V**:
	xxxxxxx		B_5		xxxxxxx				"Violence
	xxxxxxx				xxxxxxxxx				on TV"
	xxxxxxxx		S_b		xxxxxxxx			F_e	
	xxxxxxx				xxxxxxxxx				**F**:
	xxxxxxxxx				xxxxxxx				"Facebook
0	xxxxxxxx				xxxxxxxx				Privacy"
	xxxxxxxx				xxxxxx				
	xxxxx				xxxxxxx				
	xxxxxx	B_4			xxxxx		L_e V_b		
	xxxx				xxxxx		F_d		
	xxxx				xxxxx				
	xxx		G_b		xxxx				
	xxxx	B_2			xx		F_b		
	xx				xxx				
-1	xx				xx				
	xxx				xx				
	xxx	O_1			xxx				
	xx	B_3			x				
	x	O_2			xx				
	x				x				
	x	B_1			x				
	x	O_3			x				
	x	O_4					L_d		
	x				x		L_ab		
-2	x				x				
		G_a				L_c			
	x					F_c			
		S_a				V_a			
-3									

In the ability distribution of each dimension, each ‘x’ represents 3.7 students. We can first see that the mean and the variance of the ability distribution are similar between the two dimensions. For the scientific argumentation dimension, the mean is 0 logit and the variance is 0.59 logit. For the general argumentation dimension, the mean is 0.01 logit and the variance is 0.61 logit. Furthermore, the range of the item threshold locations covers the ability distribution well in each dimension.

A main finding that can be seen in Figure 4 is that the 3-level structure of our hypothesized learning progression is supported across the two different contexts. Items designed to test Level 0 of the hypothesized map tested, on average, the easiest. Items designed to test Level 1 were of intermediate difficulty. Items designed to test Level 2 tested were the most difficult.

The second research question addressed in this paper asked, “To what extent is student competence with increasingly complex argumentation generalizable?” In other words, how specific are these findings to *scientific* argumentation? The Wright maps depicting item thresholds in Figure 4 offer evidence that there may be something particular about argumentation situated in scientific contexts that makes it more difficult than argumentation in familiar contexts. For instance, Figure 4 shows that Level 1 tasks, or tasks that required students to construct warrants and complete arguments, were easier when situated in the general argumentation context. Level 1 item thresholds were, on average, lower for the general argumentation tasks than the tasks that relied on knowledge of the particulate structure of matter. This could be due to the fact that students need specific content knowledge in order to construct the warrants between claims and evidence in a scientific scenario. Hence students with formal educational exposure to the physical process of dissolving may find it easier to construct an argument that the sugar remains, despite no longer being visible because they have a model of the underlying physical process. However, the reader should bear in mind that while a reliance on specific scientific content knowledge is the most likely explanation of the elevated difficulty, these findings do not eliminate other possibilities.

However, Figure 4 shows that Level 2 tasks, which required both comparison and critique, were very difficult for eighth grade students, regardless of the context. Even the tasks situated in familiar contexts were high on the difficulty scale. This points to some domain-general difficulty in coordinating multiple claims, evidence, and warrants. Such tasks appear to be cognitively difficult for middle school students. We conducted follow-up work in which a similar set of items was administered to a sample consisting of both eighth grade *and* tenth grade students (159 tenth-graders and 119 eighth-graders in the same large, urban school district in the Bay Area) in 2014. These students took three complete item bundles consisting of 19 total items. These argumentation items were all situated in a scientific context. The results showed that older students had higher ability with argumentation, in general, and were able to complete Level 2 tasks more readily. Figure 5 shows the distribution of argumentation ability across eighth grade students and tenth grade students in this study. The tenth-graders were clustered near the top of the ability distribution. The eighth graders possessed a range of ability; however; on average, had lower argumentation ability than the tenth-graders.



Based on the follow-up work, we believe it is possible that if the initial work had been conducted with higher ability students, there would have been a more detectable difference in the difficulties of Level 2 tasks across scientific and general domains.

Discussion

This study provides empirical evidence for a learning progression describing students' facility with argumentation. The lowest level of the learning progression describes the advancing of a claim, and the competency to identify of claims, warrants and evidence—these are tasks that do not require the coordination of any elements of argument. The next level of the learning progression describes the coordination of at least two elements of an argument. Such ability is dependent on the facility to identify the elements of an argument and then explain the relationship between them, which increases the intrinsic cognitive load on working memory. The top-most level of the map describes comparison and critique, activities that require the coordination of multiple elements of an argument. The three-level structure of progression was supported by unidimensional and multidimensional Rasch analysis. Item parameters from Rasch analysis modeling also suggested that tasks involving critique were more difficult for students, a conclusion which was supported by findings from cognitive interviews. Thus, the hypothesized progression was altered to reflect these findings.

Implications for curriculum

Researchers disagree about the degree to which curriculum and instruction should be linked to learning progressions. Duncan and Hmelo-Silver (2009) contend that learning progressions are not developmentally inevitable, but rather dependent on the instruction that students receive. That is, curriculum and instruction inevitably affect learning trajectories. However, they also point out that learning progressions are "theoretical constructs that are not intended to be tied to specific instructional interventions" (p. 608). Research studies in the past have differed in the emphasis placed on instruction and whether or not instructional interventions are seen as necessary for empirical validation of progression. For example, Schwarz et al. (2009) and Songer et al. (2009) describe extensive instructional interventions and view these as critical for the evaluation of the validity of any learning progression. On the other hand, Mohan, Chen and Anderson (2009) and Duncan, Rogat and Yarden (2009) place less emphasis on instructional interventions and argue that development and validation can be informed by what students know and are able to do under the status quo conditions of their education. The work presented here aligns more closely with the latter studies. Consequently, there is first an open question of how student performance might change with instruction? In short, would the mean level of performance be significantly better if argument and critique became a more explicit feature of the science curriculum? Given the absence of argument both from texts (Penney, Norris, Phillips, & Clark, 2003) and the classroom (Newton, Driver, & Osborne, 1999; Weiss & Pasley, 2004), our view is that it could only improve.

Second, there is the question of how this progression might inform future instruction of argumentation? In short, what do our results say about the teaching and learning of argumentation? In answer, our validated progression offers a sequence with which to introduce argumentation, a practice that is unfamiliar to many students. As Corcoran, Rogat and Mosher (2009) point out, learning progressions are based on evidence from real students about what they know and can do, and thus developing curriculum around learning progressions is more theoretically justified than other methods, such as a presumed "logical" ordering of topics or a historically-determined curriculum. The learning progression for argumentation presented in this paper suggests that middle school students are capable of identifying and making claims, selecting evidence to support a claim, and even providing reasoning that links claim and evidence. Furthermore, it is quite possible that they are capable of higher level argumentation; however, completing written tasks that demonstrate this competency appears to be very challenging for most eighth-graders. Thus, it would be reasonable for curriculum in the middle grades to focus on the foundations of argumentation, helping students to develop a confidence with the notion of claim, evidence and argument and introduce higher level tasks (such as comparison and critique) in a structured and scaffolded way. For instance, tasks could ask students to identify flaws in an experimental design or a conclusion drawn from the interpretation of a set of data.

Moreover, findings from this work suggest that critique is relatively difficult for students. This may be a function of the way in which science is taught typically, as a series of facts or concepts to be understood, rather than a set a set of idea to be *argued for* (Weiss, Pasley, Smith, Banilower, Heck, 2003). Students are not often asked to construct a scientific explanation and justify it with an argument. Moreover, the teachers who collaborated on this research reported that critique is an unfamiliar task for eighth grade students. Students are rarely, if ever, expected to identify what the flaws are in a model, to explain the weaknesses in an experimental design, or identify the failings in an interpretation of a given data set. A curricular implication of this work, then, is that students do need support to engage in critique - perhaps by using sentence starters such as “a weakness in the argument is…” in order to develop this facility. Perhaps more fundamentally, if we expect students to be able to mount critiques, this practice should occupy a more prominent role in school science (Henderson et al., 2015).

Furthermore, results suggest differences in student competency with argumentation depending on whether it is presented in a scientific context or in a general context. Our findings suggest that the level of scientific content knowledge does affect students’ ability to argue, since tasks situated in general scenarios were easier for grade 8 students – a finding which concurs with other work (Osborne et al., 2004; Von Aufschneiter et al., 2008. In particular, the *construction* of scientific arguments was more difficult than the *construction* of general arguments for students. However, *critique* of arguments was very difficult for students, regardless of context. An implication of this finding is that argumentation needs to be taught in a discipline-specific manner—that is in the science class, rather than being taught only in humanities with the assumption that there will be transfer to science.

Implications for assessment

The work reported here has perhaps the greatest implications for the assessment of science achievement. An important product of this work is a set of valid, reliable items to assess scientific argumentation and a list of "lessons learned" from the process of development. Most importantly, learning progressions, or the construct maps that are linked together to form learning progressions, offer an explicit model of cognition that can underlie the design of assessments (Wilson, 2013). Without a learning progression, assessment design operates under an implicit model of cognition and a lack of guidance for item development that have high construct validity. The learning progression presented here offers a way in which to operationalize argumentation in written assessments that can be administered to individual students for the purpose of formative or summative assessment. As a case in point, teachers and researchers can use the argumentation diagrams in our progress map (Table 1) to estimate the relative complexity of argumentation assessments by identifying which of the shaded elements students are being asked to explicate on assessments. And without carefully accounting for what assessments specifically ask for, we do not know whether students were unable to engage in certain aspects of argumentation, or whether it was not explicitly asked for.

In addition to providing a progress map that is based on a model of cognition for argumentation, this work prompted the development of specific principles of assessment design for argumentation. We found that increasing the number of claims and pieces of evidence a student has to juggle within a task increases the difficulty of the task. For example, an early version of the "Sugar in Water" item contained 3 different claims and 6 pieces of evidence. The Level 0 items in that bundle were as difficult as some of the Level 2 items in bundles on the same assessment because of the demands made on cognitive load. Thus, in future iterations, we held the number of claims in each bundle constant (2).

The role of scaffolding in an assessment item became important and deserves further attention in future work. In our work, it appeared that eighth graders benefited from scaffolding, in the form of sentence starters, to complete argumentation tasks. For example, earlier versions of "critique" items did not provide the specific sentence starter, "The problem with this argument is . . ." and the majority of students defaulted to simply providing an alternative to the argument, rather than a specific critique. This was problematic, since providing an alternative argument is a socially acceptable argumentative move, but such a response does not provide evidence about whether a student is capable of engaging in critique. Furthermore, until argumentation has become a standard element in classrooms, students may be unfamiliar with terms such as "claim," "evidence" and certainly concepts such as "warrant" or "rebuttal." Thus, designing assessments that include sentence starters, especially for higher-level argumentation tasks, increases the likelihood that students will enter the correct problem space and provide more reliable evidence of their competency with argumentation.

Findings from multiple rounds of item analysis during this project revealed that multiple choice items testing argumentation do not elicit the same competencies as constructed response items, and the difficulty of multiple choice items appears artificially low on Wright Maps. Multiple-choice items are attractive assessment options, since they can be scored reliably and cheaply. However, in earlier iterations of the assessments, both multiple choice and constructed response items were developed, and the multiple choice items were frequently mis-fitting and/or artificially easy. A question for the development of future items is what can be preserved from our attempt to develop selected response items, and be re-formatted to take advantage of computers such that this research is really applicable to large-scale testing?

Limitations

This study presents a hypothesized learning progression for argumentation that presents both broad levels of argumentation (i.e., Level 0, Level 1 and Level 2) and more fine-grained sublevels within each of these three main levels (e.g., 0a, 1b, 2c, etc.). While the broad levels of argumentation maintained acceptable separation on the Wright Map after accounting for 99% confidence intervals, the sublevels were not consistently ordered in each set of analyses. As there were many fewer items testing each sublevel than there were items testing the three main levels, the focus of this study was on how the data validated our hypothesized learning progression more broadly. Decreasing the noise in the sublevel ordering requires a more powerful future study. We also did not directly test the effect of context on argumentation (scientific versus general); rather, we relied on evidence of relative difficulty. It would require a treatment-based study to really investigate the effect of context—such work would be a great candidate for future study.

An additional limitation of this work is that the learning progression and accompanying assessment items do not describe or measure argumentation as a social practice, as students did not engage in a dialogue while completing these items. Rather, what the progression measures is students' ability to engage individually in argumentative reasoning in both a scientific and more general context. However, we believe that this limitation does not pose a serious threat to validity, and we offer two reasons for this. First, findings from think aloud interviews suggest that students do engage in a simulated dialogue while completing the tasks, especially near the end when they are weighing multiple arguments. When they are asked which argument they agree with more, many students were able to verbally describe how they were weighing both options, and ultimately constructing a defense of one and a critique of another. In that sense, argumentation occurs because students are able to engage in a dialogue intra-mentally (Billig, 1996). Second, argumentation as practiced by scientists often occurs exclusively in writing, via published articles and written responses. Written arguments are often the foundation of the practice of argumentation; hence these types of assessments measuring students' facility with written argumentation reflect the way in which argumentation is undertaken in the scientific community. Argument is a product while argumentation is process, and we acknowledge that this study focused exclusively on the product.

A final limitation is the narrow focus on cognitive progression without consideration of cultural and social-emotional factors that may affect students' measured ability on the argumentation construct. Scholars have offered observations about the ways in which different cultures view argumentation. For example, Becker (1986) argues that argumentation is not a normal feature of East Asian (Chinese and Japanese) discourse, pointing out that in these cultures age is equated with authority such that students questioning a master (teacher) would be considered unacceptable. Furthermore, taking an adversarial position would be seen as occupying a position of personal rivalry and antagonism. Such cultural barriers would possibly hinder students engagement in productive argumentation. The progression presented in this paper does not account for reluctance due to cultural actors. To our knowledge, no studies exist that empirically test whether students in different cultures experience cultural barriers to argumentation. However, researchers have pointed out that argumentation is often an unfamiliar practice and students who are not used to engaging in questioning, critique or argument, such a practice would undermine or threaten their need for psychological safety. Nasir, Roseberry, and Warren (2006) propose to make it less threatening by making the structure of the domain visible—in thise case, discussing the nature of an argument, its purpose, and to find ways in which the individual can be separated from the ideas. In this work, there were limited opportunities to support students, who were likely unfamiliar with the practice of argumentation, by making the practice more visible. Additional work could shed light on how interventions like this might modulate students' measured ability on the argumentation spectrum.

Future directions

Future directions for this work fall into three broad categories: (1) further development and description of the learning progression presented here, (2) design of assessments, and (3) instructional interventions designed to leverage the learning progression and accompanying assessments.

The limitations of the current study that are worthy of attention in future work are further elaboration of the sublevels of the map and additional investigation of the highest level of the map. Reliable ordering of the sublevels of the map may not be possible; however, further investigation, either through a more powerful study (more students in the sample) or a detailed qualitative study examining interviews and constructed responses from students, might shed light on how the smaller sublevels build toward mastery of a broad level of argumentation. Investigation of the highest levels of argumentation would require a study with older students. We reported in the findings that work with 10th graders showed that they had higher ability on the argumentation construct than the 8th graders used in the main study. Thus, administering items to students in their final years of high school would offer more data on the highest levels of the map. Such a study would offer a “top anchor” for our learning progression, which was elusive in the study reported here.

A second category of future work is investigation of instructional interventions that utilize the argumentation learning progression or accompanying assessments. Perhaps a major question raised by this learning progression is whether student facility can be improved by appropriate instruction? To date, there have been limited attempts to explore how argumentation might be taught. Our progression would suggest that student facility is initially dependent on the ability to identify the elements of an argument. Exercises that asked students to distinguish claims, warrants and data would provide insight into the nature of argument in science which might then provide a foundation for more demanding exercises which asked students to construct their own arguments. Engaging students in critique would ask them to think why an argument might be flawed and require students to invoke appropriate domain-specific knowledge. Direct critique of a scientific idea may be a more difficult task for students as human reasoning suffers from confirmation bias (Nickerson, 1998). Moreover, Baron (1995) has found that people were more likely to rate one-sided arguments higher than two-sided arguments, suggesting reluctance to critique may be related to perceptions of what makes an argument strong. Yet research studying the use of "refutational texts" shows that students learn science more effectively when they can understand why the wrong idea is wrong, rather than why the right idea is right (Hynd & Alverman, 1986). Thus, engaging students in critique would help to develop their competence both with the higher levels of argumentation and their understanding of the science content. As critique is not common practice in science classrooms, it will need support in the form of modeling and scaffolding if students are to reach higher levels of competence with argumentation.

Finally, future work on the design of assessments is a natural outgrowth of this work. In particular, we can imagine improvements to these assessments that leverage technology, including agent-based computer argumentation, computer-adaptive testing, and natural language processing for the automated scoring of constructed responses. We are conscious of the fact that, in their current form, our assessments are reliant on open-ended responses that require hand scoring. Given the increasing move to the computerization of assessment for the Common Core in the USA and for PISA 2015 by the OECD, there is a need to explore ways in which student progression with argumentation, as represented by our map, can be operationalized in ways that can be machine scored. What we hope we have offered in this work is a deeper understanding of what might constitute a developmental progression with argumentation and a foundation on which others can address the issues above.

References

Achieve (2013). Next Generation Science Standards. http://www.nextgenerationscience.orghttp://www.nextgenerationscience.org

Allchin, D. (2012). Teaching the nature of science through scientific errors. Science Education, 96(5), 904-926. doi: 10.1002/sce.21019

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, teaching and Assessing: A revision of Bloom's Taxonomy of Educational Objectives*. London: Longman.

Banilower, E. R., Smith, P. S., Weiss, I. R., Malzahn, K. A., Campbell, K. M., & Weis, A. M. (2013). Report of the 2012 National Survey of Science and Mathematics Education. Horizon Research, Inc.

Baron, J. (1995). Myside bias in thinking about abortion. Thinking & Reasoning, 1(3), 221–235. http://doi.org/10.1080/13546789508256909

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. doi:10.1002/sce.20402

Billig, M. (1996). Arguing and Thinking: A Rhetorical Approach to Social Psychology. Cambridge University Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company.

Bruner, J. S. (2009). The Process of Education, Revised Edition. Harvard University Press.

Cavagnetto, A. R. (2010). Argument to Foster Scientific Literacy A Review of Argument Interventions in K–12 Science Contexts. Review of Educational Research, 80(3), 336–371. doi:10.3102/0034654310376953

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. Cognition and instruction, 8(4), 293–332.

Chang, Y-H, Chang, C-Y, & Tseng, Y-H (2010). Trends of Science Education Research: An Automatic Content Analysis. Journal of Science Education Technology, 19, 315-331.

Corcoran, T., Mosher, F. A. & Rogat, A. (2009). Learning Progressions in Science: An evidence-based approach to reform. Philadelphia, PA: Consortium for Policy Research in Education.

Crombie, A. C. (1994). Styles of scientific thinking in the European tradition: the history of argument and explanation especially in the mathematical and biomedical sciences and arts. 2 (1994). Duckworth, London.

Disessa, A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), Constructivism in the computer age (pp. 49–70). Lawrence Erlbaum.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. Science education, 84(3), 287-312.

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. Journal of Research in Science Teaching, 46(6), 606–609. doi:10.1002/tea.20316

Duncan, R. G., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening students' understandings of modern genetics across the 5th–10th grades. Journal of Research in Science Teaching, 46(6), 655–674. doi:10.1002/tea.20312

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. Science Education, 88(6), 915–933. doi:10.1002/sce.20012

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data (revised edition). Cambridge, MA: MIT Press.

Fischer, Frank , Kollara, Ingo , Uferb, Stefan , Sodiana, Beate , Hussmannc, Heinrich , Pekruna, Reinhard , . . . Eberlea, Julia (2014). Scientific Reasoning and Argumentation: Advancing an Interdisciplinary Research Agenda in Education. *Frontline Learning Research, 5*, 28-45.

Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. Science Education, 92(3), 404–423. doi:10.1002/sce.20263

Gagne, R. M. (1968). Learning hierarchies. Educational Psychologist, 6, 1–9.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. Science Education, 94(2), 259–281.

Gotwals, A. W., & Songer, N. B. (2013). Validity Evidence for Learning Progression-Based Assessment Items That Fuse Core Disciplinary Ideas and Science Practices. Journal of Research in Science Teaching, n/a–n/a. doi:10.1002/tea.21083

Gotwals, A. W., Songer, N. B., & Bullard, L. (2012). Assessing students' progressing abilities to construct scientific explanations. In *Learning Progressions in Science* (pp. 183-210). SensePublishers.

Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. Journal of Research in Science Teaching, 28(9), 799–822. doi:10.1002/tea.3660280907

Hannaway, J., & Hamilton, L. (2008). Performance-based accountability policies: Implications for school and classroom practices. The Urban Institute and RAND Corporation.

Harrison, A.G., & Treagust, D.F. (2000). Learning about atoms, molecules, and chemical bonds: A case study of multiple-model use in grade 11 chemistry. Science Education, 84(3), 352–381.

Henderson, B., Osborne, J., MacPherson, A., & Wild, A. (2015). Beyond Construction: Five arguments for the role of critique in teaching science. *International Journal of Science Education, 37*(10), 1668-1697.

Howson, Colin, & Urbach, Peter. (2006). *Scientific Reasoning: A Bayesian Approach* (3rd ed.). Chicago: Open Court.

Hynd, C., & Alvermann, D.E. (1986). The Role of Refutation Text in Overcoming Difficulty with Science Concepts. Journal of Reading, 29(5), 440–46.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. Zeitschrift Für Psychologie / Journal of Psychology, 216(2), 61–73. doi:10.1027/0044-3409.216.2.61

Koslowski, B., Marasia, J., Chelenza, M., & Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cognitive Development, 23*(4), 472-487.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory into practice, 41(4), 212-218.

Kuhn, D. (1999). A Developmental Model of Critical Thinking. *Educational Researcher, 28*(2), 16-46.

Kuhn, D., Wang, Y., & Li, H. (2011). Why Argue? Developing Understanding of the Purposes and Values of Argumentive Discourse. *Discourse Prcesses, 48*, 26-49.

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. Journal of Research in Science Teaching, 51(5), 581–605. doi:10.1002/tea.21147

Lee, M., Wu, Y., & Tsai, C. (2009). Research Trends in Science Education from 2003 to 2007: A content analysis of publications in selected journals. International Journal of Science Education, 31(15), 1999–2020. doi:10.1080/09500690802314876

Lehrer, R., & Schauble, L. (2006). Scientific Thinking and Science Literacy. Handbook of Child Psychology (pp. 153-196). New York: Wiley.

Lemke, J. L. (1990). Talking science: Language, learning, and values. Norwood, NJ: Ablex.

Longino, H. E. (1990). Science as Social Knowledge: Values and Objectivity in Scientific Inquiry. Princeton University Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149–174. doi:10.1007/BF02296272

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. Behavioral and brain sciences, 34(02), 57-74.

Messick, S. (1987). Assessment in the schools: Purposes and consequences. Educational Testing Service.

Messick, S. (1991). Psychology and methodology of response styles. Improving inquiry in social science: A volume in honor of Lee J. Cronbach, 161–200.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81-97.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. Educational Measurement: Issues and Practice, 25(4), 6-20.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. Journal of Research in Science Teaching, 46(6), 675–698. doi:10.1002/tea.20314

National Research Council. (2007). Taking Science to School: Learning and Teaching Science in Grades K-8. National Academies Press.

National Research Council. (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC.: Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education.

Newton, P., Driver, Rosalind, & Osborne, J.F. (1999). The Place of Argumentation in the Pedagogy of School Science. International Journal of Science Education, 21(5), 553-576.

Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220. http://doi.org/10.1037/1089-2680.2.2.175

OECD. (2012). OECD Science, Technology and Industry Outlook 2012. Paris: Organisation for Economic Co-operation and Development.

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. doi:10.1002/tea.20035

Osborne, J. (2010). Arguing to Learn in Science: The Role of Collaborative, Critical Discourse. Science, 328(5977), 463–466. doi:10.1126/science.1183944

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education, 95*(4), 627-638. doi: 10.1002/sce.20438

Osterlind, S. J. (1998). Constructing Test Items: Multiple-Choice, Constructed-Response, Performance and Other Formats. Springer Science & Business Media.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. Educational psychologist, 38(1), 1–4.

Penney, K., Norris, S. P., Phillips, L., & Clark, G. (2003). The Anatomy of High School Science Textbooks. *Canadian Journal of Science and Mathematics Education, 3/4*, 415-436.

Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. Learning and Instruction, 12(1), 61–86. doi:10.1016/S0959-4752(01)00016-0

Rasch, G. (1966). An Item Analysis Which Takes Individual Differences into Account. British Journal of Mathematical and Statistical Psychology, 19(1), 49–57. doi:10.1111/j.2044-8317.1966.tb00354.x

Rosenbaum, P. R. (1988). Items bundles. Psychometrika, 53(3), 349-359.

Schwartz, R., & Ayers, E. (2011). Delta dimensional alignment: Comparing performances across dimensions of the learning progression for assessing data modeling and statistical reasoning. Unpublished manuscript, University of California, Berkeley, Berkeley, CA.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., … Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. Journal of Research in Science Teaching, 46(6), 632–654. doi:10.1002/tea.20311

Shavelson, R. J. (2009). Reflections on Learning Progressions. Paper Presented at the Learning Progressions in Science Conference, Iowa City, IA.

Shavelson, R. J., & Kurpius, A. (2012). Reflections on learning progressions. In*Learning progressions in science* (pp. 13-26). SensePublishers.

Songer, N. B., Kelcey, B., & Gotwals, A.W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. Journal of Research in Science Teaching, 46(6), 610–631.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. Learning and instruction, 4(4), 295–312.

Sztajn, P., Confrey, J., Wilson, P. H., & Edgington, C. (2012). Learning trajectory based instruction toward a theory of teaching. Educational Researcher, 41(5), 147–156.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. Journal of Research in Science Teaching, 46(6), 699–715. http://doi.org/10.1002/tea.20308

Thagard, P. (2008). Explanatory coherence. In Adler, J. and Rips, L. (Eds), Reasoning: Studies of human inference and its foundations (pp. 471-513). Cambridge: Cambridge University Press.

Toulmin, S. (1958). The Uses of Argument. Cambridge: Cambridge University Press.

Tricot, André, & Sweller, John. (2013). Domain-Specific Knowledge and why Teaching Generic Skills does not Work. *Educational Psychology Review*.

Van Eemeren, F. H., & Grootendorst, R. (2004). A Systematic Theory of Argumentation: The Pragma-dialectical Approach. Cambridge University Press.

Von Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. Journal of Research in Science Teaching, 45(1), 101–131.

Walton, D. N. (1990). What is Reasoning? What Is an Argument? The Journal of Philosophy, 87(8), 399–419. doi:10.2307/2026735

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*: Harvard University Press.

Weiss, I. R., Pasley, J. D, Sean Smith, P, Banilower, E. R., & Heck, D. J. (2003). A Study of K–12 Mathematics and Science Education in the United States. Chapel Hill, NC.: Horizon Research.

Wiliam, D. (2010). What Counts as Evidence of Educational Achievement? The Role of Constructs in the Pursuit of Equity in Assessment. Review of Research in Education, 34(1), 254–284. doi:10.3102/0091732X09351544

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. Journal of Research in Science Teaching, 46(6), 716–730. doi:10.1002/tea.20318

Wilson, M. (2013). Constructing Measures: An Item Response Modeling Approach. Routledge.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. Applied Measurement in Education, 13(2), 181-208.

Wright, B. D., & Masters, G. N. (1982). Rating Scale Analysis. Rasch Measurement. MESA Press, 5835 S. Kimbark Avenue, Chicago, IL 60637 ($25). Fax: 773-702-1596; Fax: 773-834-0326; e-mail: MESA@uchicago.edu; Web site: <http://www.rasch.org>. Retrieved from http://eric.ed.gov/?id=ED436551

Wu, M. L., Adams, R. J., & Wilson, M. (2012). ConQuest: Generalized item response modelling software (Version 3.0). ACER.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172-223.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. Journal of Research in Science Teaching, 39(1), 35–62. doi:10.1002/tea.10008