A NEW LEARNING PROGRESSION FOR STUDENT ARGUMENTATION IN SCIENTIFIC CONTEXTS

J. Bryan Henderson, Jonathan Osborne, Anna MacPherson and Evan Szu
Stanford University, Graduate School of Education, Stanford, California, USA

Abstract: Learning progressions offer a model for educators of the paths by which learning might proceed within a domain. A particular emphasis of much research – both within science education and other domains – has been on teaching higher order thinking skills and engaging students in evidence-based argumentation. However, no tested and validated model exists of students’ progression with argumentation in science education. We present the theoretical underpinnings of a model for argumentation based in the work of Toulmin that has been developed and tested over two annual cycles. In each cycle, working with a large sample (n=600) of 14-year-old students, student think-aloud interviews and item co-development with their teachers has allowed us to monitor construct and response process validity. This work is important for developing a body of expertise in how to assess argumentation in the context of science and for establishing an empirically tested model of the nature of student progression with this competency.

Keywords: argumentation, learning progressions, assessment, science education

BACKGROUND AND FRAMEWORK

The value of learning progressions in science

Commonly, the goals of science education are defined in terms of a hierarchy of knowledge that is intrinsic to the discipline. A newer approach on how to structure science education has emerged in recent years from the National Research Council (2007) report Taking Science to School. This argues for maps of “successively more sophisticated ways of learning about a topic that can follow and build on one another as children learn and investigate a topic over a broad span of time.” Such maps, commonly called learning progressions, outline potential cognitive paths that students might follow as they develop a more sophisticated understanding of a core scientific concept. Thus learning progressions can both serve as a valuable diagnostic tool for teachers of science content in that domain, while also informing policy makers on how to better align science content standards with what students can realistically achieve.

The value of argumentation in science

Despite extensive research literature suggesting that argumentation supports student conceptual understanding (Osborne, 2010), what happens in science classrooms is largely motivated by factors such as teachers’ previously held beliefs, state science content standards, and the format and content of state science assessments (Banilower et al., 2013). If reasoning and arguing are to be emphasized in science classrooms, such higher-order skills must be represented both in the standards and on the tests. However, little is known about how to readily assess reasoning and argumentation in a
domain-specific manner, particularly in a quantitative or semi-quantitative way. Investigating scientific argumentation and reasoning through the lens of a learning progression provides just such a way to operationalize and empirically test these important science skills.

**Previous research**

While Schwarz et al. (2009) have developed a learning progression for scientific modeling that contains explanatory elements, Berland and McNeill (2010) propose the only learning progression to date that specifically focuses on scientific argumentation. In their progression, the authors use examples from elementary, middle, and high school science classrooms to characterize 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 some preliminary evidence as to how argumentation ability develops over time, as they found the structure of argumentation was more complex in 12th grade students’ work than that of their middle school counterparts.

The progress map Berland and McNeil (2010) propose is based on qualitative observations of the sophistication and frequency of various forms of argumentation and consists of three dimensions – *instructional context, argument product, and argument process*. This suggests that multiple progress variables would need to be operationalized and measured to validate their learning progression. What their map lacks is any systematic attempt to empirically probe student understanding and measure competency with argumentation in individual students. Hence it is difficult to say whether students were unable to engage in certain aspects of argumentation, or whether they simply chose not to because it was not specifically asked for. Thus, while Berland and McNeil clearly and thoroughly delineate three plausible dimensions of scientific argumentation using specific classroom examples that exemplify each dimension, their progress map lacks empirical validation.

In contrast, we propose a learning progression for argumentation that is amenable to empirical validation using assessment items that target different levels of our progress map. We have designed these assessment items to elicit student responses commensurate with specific levels of competency as described by our learning progression. Data generated from student responses provides information about the extent to which these responses support our hypothesized progression. Lack of empirical support can be attributed to a lack of validity of the assessment items, a lack of validity of the proposed learning progression, or possibly both. Therefore, an iterative process that holistically examines both the assessment items and the progress map has enabled us to develop an empirically validated learning progression for argumentation. It should be cautioned, however, that validation of such a progression does not entail that every single student moves through the progression in the same manor (Duncan & Hmelo-Silver, 2009). Rather, validation of our proposed learning progression for argumentation can provide evidence that an understanding of how to argue in a scientific context may, on average, develop to some degree in accordance with our hypothesized map.
OUR PROPOSED LEARNING PROGRESSION FOR ARGUMENTATION

Towards a working definition of argumentation

We view argumentation as not simply an aptitude that can be assessed, but rather, as a competency. Koeppen et al. (2008) define competence as “context-specific cognitive dispositions that are acquired and needed to successfully cope with certain situations or tasks in specific domains” (p. 62). Building on this conception of competency, we view scientific argumentation as a complex form of reasoning demanded by situations that require scientific content knowledge to construct and/or critique proposed links between hypotheses and evidence. By invoking the term reasoning, we draw on Walton’s (1990) description of theoretical reasoning, which he characterizes as a discursive process “oriented to finding reasons for accepting a proposition as true or false” (p. 405). Note that Walton does not see reasoning and argumentation as equivalent, but rather, argumentation is a specific form of reasoning that seeks resolution to one or more claims:

- Argument is a social and verbal means of trying to resolve, or at least contend with, a conflict or difference that has arisen or exists between two or more parties. In an asymmetrical case, one party puts forward a claim, and the other party questions it. In a symmetrical case, each party has a claim that clashes with the other party’s claim. The claim is very often an opinion, or claim that a view is right, but it need not be (p. 411).

Given Walton’s (1990) depiction of argument as the adjudication among multiple claims, we turn to Toulmin’s (1958) model for the structure of practical or informal argument. In contrast to the logic of more traditional deductive arguments that start from first principles to ultimately arrive at a claim, practical arguments instead begin with claims and seek justification to support those claims. We see this model as more in line with how argument is used in scientific practice – scientists begin with a claim (i.e., a hypothesis) and test its veracity against empirical data. 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).

Toulmin’s model (1958) begins with a claim as a “conclusion whose merits we are seeking to establish” (p. 90). It follows that in seeking to establish the merit of a claim, one turns to data or evidence that may lend support. In turn, 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. For the purposes of our work, claims, warrants, and evidence play a key role in the development of our learning progression, the incorporation of which we will detail below.

While Toulmin’s (1958) model of practical argument plays a central role in our learning progression for argumentation, it is not sufficient. Indeed, in addition to justification, the process of critique must be also incorporated into our progress map. As Ford (2008) has
argued, the goal of scientific reasoning is knowledge construction, and this is achieved by a dialectic consisting both of construction and critique. Argument is used to justify the validity of explanatory hypotheses, experimental designs and interpretations of a given data set. Critique is essential to identifying flaws in such arguments.

The ability to engage in critique, however, requires somewhat different competencies. For instance, rather than constructing a claim, it requires the ability to identify what is the claim in any given argument, what elements constitute the data that is used to support the argument, and what kind of reason is used to relate the data to the claim. Such competency is reliant on at least a tacit meta-knowledge of the features of an argument and the ability to distinguish its component elements. Taken a step further, it requires the ability to construct a rebuttal that would explain why the reasoning in a given argument is flawed. 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. Recently, work by Evagorou and Osborne (2007) and others (Clark, Sampson, Weinberger, & Erkens, 2007) has led to a deeper theoretical understanding of the essential features of good quality argumentation.

In summary, we conceptualize scientific argumentation as a competency, and highlight several key aspects of this competency that we believe our progress map should account for:

1) Argumentation is a specific type of reasoning that seeks resolution to one or more claims (Walton, 1990).

2) When viewing argumentation as a justificatory exercise, a claim must be considered in concert with both evidence and a warrant that purports to establish the relation between the evidence and the claim (Toulmin, 1958).

3) In addition to constructing an evidentiary foundation for a claim with data, the process of critique is also essential for identifying potential flaws in argument constructions (Ford, 2008).

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. Given that opportunities to engage in argumentation rarely occur in science classrooms (Newton, Driver, & Osborne, 1999), and the multitude of information that must be processed to construct/critique a scientific argument, 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**

Given that students have little experience with scientific argumentation, any progress map for scientific argumentation should also account for limitations in human working (i.e., short-term) memory. This is because working on novel tasks places high cognitive demand on human working memory. These limitations in working memory are perhaps most famously documented by Miller (1956) who described how, across multiple cognitive experiments, the number of entities that could be simultaneously held in human
working memory is SEVEN +/-2. These findings have served as a foundation upon which cognitive load theory (Chandler & Sweller, 1991; Paas, Renkl, & Sweller, 2003; Sweller, 1994) has been built. Pollock, Chandler, and Sweller (2002) delineate four primary assumptions that cognitive load theory is predicated upon:

1) Human working memory can only process a limited amount of information at a given time, per the SEVEN +/-2 findings documented by Miller (1956).

2) Human long-term memory is essentially limitless.

3) This long-term memory can hold schemas, which “chunk” together fragmented bits of information. These schemas require less capacity when retrieved by the working memory than were all the individual bits retrieved independent of the schema tying them together.

4) Schemas that become automated through repetition also reduce working memory load through automatic processing that bypasses working memory altogether.

Furthermore, Pollock, Chandler, and Sweller (2002) summarize two basic sources of cognitive load. The first source – intrinsic cognitive load – “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, extraneous cognitive load “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, cognitive load theory offers us a lens through which the competency of scientific argumentation can be operationalized as a single progress variable. More specifically, given that Toulmin (1958) resolves argumentation into fundamental constituents such as claim, warrant, and evidence, the process of constructing or critiquing arguments can be viewed as an orchestration of various combinations of these elements of argument. It follows that for tasks that increase the number of elements that must be processed, the intrinsic cognitive load on the working memory increases, thereby making it more difficult to demonstrate argumentative competency. Hence conceptualizing the elements of argument as a source of intrinsic cognitive load guides us as we designate certain argumentative tasks to be more/less difficult than others on our proposed learning progression. With this in mind, intrinsic cognitive load also assists us in considering a realistic top anchor for our progress map, as we are mindful of developing assessment items in which the cognitive demand is reasonably within the limits of human working memory. Finally, cognitive load theory also guides our creation of assessment items to probe each level of our argumentation progress map. More specifically, as our intent is to isolate the intrinsic cognitive load demanded for increasingly complicated orchestrations of Toulmin’s elements of argument, in parallel we seek to minimize the extraneous cognitive load of how our assessments are formatted. For example, we work with teacher co-developers to flag and remove potentially confusing words, figures or images. More details about the challenges and lessons learned from our attempt to create assessments scientific argumentation in the written form are summarized in Osborne et al. (2013, in prep).

**Our proposed progress map**

Drawing on the ideas of intrinsic cognitive load required to engage in construction and critique of more/less sophisticated combinations of claims, warrants, and evidence, the
current progress map for of our hypothesized learning progression for argumentation is shown in Appendix A. To demonstrate competency on items probing the earliest levels of our progression, our assessments pose situations that require students to consider claims (denoted by a box in Appendix A) and/or evidence (denoted by a circle in Appendix A). As explained previously, we borrow from Toulmin (1958) by viewing claims and evidence as the fundamental elements of our progress map – the building blocks towards more sophisticated argumentation. Also note that some elements in Appendix A are shaded while others are not. Shaded elements reflect what students are specifically asked for in an item probing a particular level. Non-shaded elements bounded by a dotted line may be implicitly considered by students when assessed, but we cannot say for sure, as students are not asked to identify them explicitly. Indeed, as middle school students do not commonly practice argumentation in science classrooms, we chose to assess only what students were explicitly asked. Hence the nomenclature of our progress map reflects this explicit assessment approach.

In addition to our progress map in Appendix A, we wish to provide some context through specific examples of how our assessments can probe each level of our progress map. Figure 1 provides an example of how our assessment items are posed. In this 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 hypothetical claims are posed which are not in agreement. This sets the stage for students to call on scientific content knowledge to construct and/or critique proposed links between hypotheses and evidence – our working definition of scientific argumentation. Our outline of our progress map will refer to the item in Figure 1 providing an example of how to probe that given level in the scientific context of how the item is posed.

Figure 1. Sample assessment item. Examples of how items can be written to probe each level of our progress map will be based on the context of this item.

Our progress map 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 a claim and/or evidence without coordinating a logical connection between them. Hence the zero prefix for these levels denotes zero degrees of coordination.

Zero degrees of coordination

We regard constructing an explicit claim (Level 0a) as the most basic demonstration of argumentation competency, as doing so does not technically require any additional knowledge of the features of an argument. Indeed, it is possible for a student to construct a claim without actually knowing what it means for a statement to function as a claim in an argument. For example, given the situation posed in Figure 1, a Level 0a assessment would be to simply ask the student taking the assessment to write down what they think happened to the sugar. Indeed, it is possible for the student to say that they think the sugar remains or vanishes without knowing their declaration functions as a claim in an argument that demands evidentiary justification. Level 0b is more advanced, as it requires the capability to identify what is the claim in any argument, and in so doing a student has to now differentiate a claim from other features (e.g., warrants and evidence) of the argument they are asked to critique. For example, a Level 0a assessment with respect to Figure 1 would ask for explicit identification of what Laura is claiming (i.e., the sugar is gone). Level 0c builds on Level 0b, as it requires the ability to identify evidence that may support a claim. For example, a Level 0c assessment with respect to Figure 1 would ask for explicit identification of evidence that supports Laura (i.e., the sugar disappears). While items probing this level only explicitly ask for a supporting piece of evidence to be identified, one cannot simply say something is “supporting” without a referent as to what is being supported. Hence both claim and warrant need to at least be implicitly considered when identifying a piece of evidence as “supporting.” Therefore we see at Level 0c the first inclusion of nomenclature representing a warrant (denoted by a box with connecting arrows in Appendix A), albeit not shaded in, as items probing this level do not ask for warrants to be explicitly constructed/critiqued.

One degree of coordination

Items that indeed require explanation of warrants mark the transition from Level 0 to Level 1 on our progress map. 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 builds on Level 0 of the progress map, 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. Assessments probing Level 1a explicitly ask for a warrant to be constructed, or alternatively, critique why a warrant advanced by someone else might be flawed. For example, a Level 1a assessment with respect to Figure 1 would ask for a possible reason Laura might give for why she thinks the sugar is gone (e.g., if the sugar was still there, we would still be able to see it). Assessments probing level 1b and 1c are more demanding, as they ask students to explicitly construct or critique a claim, evidence, and warrant simultaneously. In terms of Figure 1, a Level 1b assessment would ask for a student to state explicitly a complete argument for Laura (e.g., Laura thinks the sugar is
gone, because if the sugar remained it could be still be seen, but there is evidence that it disappears after stirring). What discriminates between these two levels is that Level 1c items ask students to offer a combination of claim/warrant/evidence that is superior to what was constructed/critiqued in an item probing Level 1b. For example, a Level 1c assessment with respect to Figure 1 would ask a student to think of a possible rebuttal Mary could make to Laura’s argument (e.g., yes the sugar disappears after stirring, but due to the conservation of matter it cannot be destroyed, so perhaps the sugar is still there but simply in a different form).

**Two or more degrees of coordination**

The most advanced levels of our progress map are prefixed with the number two to denote two or more degrees of coordination. Level 2 items require students to explicate two or more warrants. Assessments probing Level 2a require students to consider two arguments and then offer a warrant for why one argument is superior to the other. These tasks demand that a student explicitly identify the warrant used in the argument they agree with in addition to advancing a warrant for why they consider that argument to be superior. For example, a Level 2a assessment with respect to Figure 1 would ask for a student to take a side (i.e., Laura or Mary’s position) and provide explicit rationale for why they take that side in the argument (e.g., I side with Mary because the sugar simply dissolved, which means that it is still there). Level 2b builds upon Level 2a as items probing this level now ask students not only to explicate why they side with one argument, but additionally, explicitly why they found the rival argument to be inferior. For example, a Level 2b assessment with respect to Figure 1 would ask for a student to not only take a side in the argument (i.e., Laura or Mary’s position), but also to explain why they found the other argument to be inferior (e.g., I side with Mary because the sugar simply dissolved, which means that it is still there. This is a better argument than Laura’s which thinks the sugar is gone, but I know matter just can’t disappear). This requires explication of four warrants – identification of the warrant used in each of the two competing arguments in addition to a third warrant for why one argument is superior and a fourth warrant for why the other argument is inferior.

**The cognitive demands of moving along the learning progression**

Indeed, the degrees of coordination necessary to resolve and evaluate any argument multiply rapidly when students are asked to make multiple comparisons. Level 2c of our learning progression exhibits nomenclature mapping the multiple warrants necessary for a student to evaluate the various degrees of importance multiple pieces of evidence may have for a specific claim. For every additional piece of evidence that a student is asked to consider vis-à-vis a claim, an additional warrant linking that new piece of evidence to the claim in question is necessary. With only several pieces of evidence posed to a student in an assessment item, the limits of their working memory are quickly being pushed to the limit. For example, a Level 2c assessment item with respect to Figure 1 would ask for evaluation of the relative significance of each of the three pieces of evidence with respect to Laura’s claim (e.g., only the first piece of evidence where the sugar disappears supports Laura, as the other two pieces of evidence both suggest that something is left behind despite no longer being visible). In pilot work, items that presented more than three pieces of evidence proved challenging as each piece of
evidence that must be considered relative to a claim represents an additional element of intrinsic cognitive load.

To this point, Level 2d of our progress map exemplifies how quickly an argumentation task can push the limits of working memory. Here, assessment items ask students to not only adjudicate between two arguments, but also do so via construction of a third argument that they believe is superior to the previous two. Here, the nomenclature in our progress map shows no fewer than five warrants that must be explicited to demonstrate full competency at this level. Based on Figure 1, an item could be written where Laura’s claim (i.e., the sugar is gone) remains the same while Mary’s claim is made to be less decisive (e.g., the sugar may still be there, or it may be gone). Then, given the same three pieces of evidence provided, a Level 2d item would ask a student to explicitly identify an alternative explanatory hypothesis that is superior to both Laura and Mary (e.g., I think the sugar remains, because based on Evidence #2 and #3 in combination with the law of conservation of matter, the sugar cannot be destroyed and hence simply changed form. This is inconsistent with both Laura and Mary’s arguments, as Laura explicitly says the sugar is gone while Mary leaves open the possibility of the sugar being destroyed.). Level 2d, which requires competency at all previous levels, currently occupies the top anchor of our learning progression for argumentation, but empirical testing may prove that this top anchor is too ambitious for students to demonstrate competency with little prior practice with argumentation. Indeed, the complexity depicted in Appendix A suggests that Level 2 tasks in general may be too much for students to handle if they are unfamiliar or unused to scientific argumentation. Fortunately our progress map lends itself to empirical testing of this very question.

**METHODODOLOGY**

The development and validation of our argumentation questions and learning progression occurred as part of a larger 4-year assessment project of middle school science. The goal of this larger project was the investigation of how students understand and engage in argument about the structure of matter.

**Sample**

This study was conducted with middle school students from an urban school district in the western United States. The district student body was 35% Asian, 23% Latino, 11% White, 11% African American, 7% South/Pacific Islander, 10% Other Non-White, and 4% declined to state. Furthermore, 61% of students were eligible to receive free or reduced lunch and 27% of the students were designated as English Language Learners (ELLs). Sampling was conducted in six to eight classrooms of students (approximately 35 students per classroom), which provided a sample size of 210-280 per test form.

The approach to assessment development utilized in this project was based on the BEAR Assessment System (Wilson & Sloane, 2000). This approach to developmental assessment provides an integrated means of developing assessments aligned meaningfully with curricular goals and instructional activities. The developmental perspective of the BEAR Assessment System is reflected in two key ways. The first is that assessment
development is centered on *progress variables*, which are the “big ideas” around which a curriculum is structured. A progress variable is an achievement continuum defined operationally by the assessment tasks to which students respond, and that can be used to track student progress over time (Masters, Adams, & Wilson, 1990). These progress variables each represent an important set of learning goals of the curriculum which are assessed repeatedly and for which teachers will wish to have summary information at critical points during the school year. Part of the work of this project is to extend and combine progress variables into a map of progression in argumentation to provide a developmental perspective of how students’ abilities to reason scientifically may progress. Information drawn from these assessments can then be used formatively to inform decisions about student progress and the next steps in instruction. Progress variables themselves are based on comprehensive reviews of the theoretical and research background, and their usefulness is verified through empirical analyses.

A second way the developmental perspective is reflected is in the operationalization of progress variables in the form of grading rubrics for open-ended items, along with exemplars of student performance at developmentally important levels. These grading rubrics should ideally lead multiple human raters to award a similar score when independently examining the same sample of student work. The BEAR Assessment System operationalizes the four principles outlined above though four primary building blocks: the construct map, items design, outcome spaces, and the measurement model (Wilson, 2005):

1) **The construct map** represents the latent construct being probed. In the BEAR system, the developmental perspective emphasizes how students progress from less to greater expertise in the domain of interest (rather than using assessment only to measuring correctness after learning activities are completed). An essential tension when choosing construct maps is the tradeoff between coverage, which drives the creation of many construct maps representing every learning goal, and usability which limits the number of construct maps that can realistically be learned and implemented by students and teachers. It is therefore important to identify the most important learning trajectories to represent as construct maps. This is usually accomplished through domain analysis that considers extant literature, the particular goals of related curricula, teachers’ expertise, input from other experts in the domain, and the theory guiding the larger learning progression.

2) **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.

3) **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.

4) 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 item response model known as the unidimensional partial credit model (Masters, 1982) and the multidimensional random coefficients multinomial logic model (MRCMLM) (Wilson & Wang, 1995). The models provide convenient and rich ways to model measures of person proficiency and item difficulty measures using...
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.

Assessment evaluation

All assessments went through a rigorous evaluation process that followed the four steps mentioned above. Using this model allows for an investigation of the measurement properties of the test by estimating an individual’s ability, the difficulty of each item, item fit statistics, and individual fit statistics. For this research, individual ability was interpreted as an individual’s proficiency in response on a continuum as represented on the progress map. Item difficulty was interpreted as the degree to which typical individuals generate or choose the correct answer to an item. This model is useful because it translates raw information (item scores and person estimates) into logits, thereby yielding item locations and individual ability estimates on a logit scale. Conducting this logistical transformation of the data frees them by placing them on a new scale free from the particulars of persons or items (Masters, 1982). This analysis was done using ConQuest software (Wu, Adams, and Wilson, 1998). The software provides visual maps and results that can be interpreted using the progress map as a guide. Discrepancies between empirical data and our hypothesized learning progression can then be illuminated by comparing the response patterns of students responding to our items with response patterns predicted by our hypothesized progress map.

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, 2005). For example, for an assessment designed to assess scientific reasoning it becomes important to verify that scientific reasoning is taking place as the student is taking the items and that he/she is not using a test-writing strategy or other means to answer questions. 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 reflected the respondent’s knowledge on achievement tests (for example). However, Messick suggests, “evidence in support of these presumptions is critical in establishing the meaning or construct validity of the scores” (Messick, 1991, p. 161). Similarly, Norris, Leighton, and Philips argue that the cognitive approaches used by students to answer questions may be difficult to predict and should be explicitly probed (2004). 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. The concurrent verbal report protocol audio-records students who are asked to complete the test while thinking aloud, occasionally with minimal prompting (e.g. “please keep talking”, “I’m listening”).

Reliability analysis

To determine if our assessments measure student understanding of scientific reasoning with sufficient consistency across individuals taking the test and across scorers assigning scores to items, we used samples of approximately 150 students (from approximately 5-7 classrooms, assuming approximately 25-30 students per classroom) per grade level. This provided an acceptable reliability estimate; Feldt (1965) suggests
that if the true reliability is .85, using a sample of 140 people will allow a 95% confidence interval for the reliability estimate to be approximately (.81, .88), and similar results may be obtained using the formulae in Feldt, Woodruff, & Salih (1987). Having 210-280 students in the sample permits stable estimates for the number of items and range of responses we are anticipating (see for example, Linacre 1994).

The Partial Credit Model (Masters, 1982) allowed us to apply a Rausch model for partial credit scoring of student responses. All items have the value of the weighted mean square statistic falling in the acceptable range of 0.75 and 1.33, suggesting the use of the Partial Credit Model is appropriate. The overall test reliability was 0.76.

Significant steps were taken to ensure inter-rater consistency. First, the scoring guide was exemplified with typical student responses using a group discussion of raters. Items were then rated using the scoring guide and ratings and assessments were then shared. Disagreements were resolved through discussion and the scoring guides amended appropriately. This cycle was continued until the differences between scorers were considered minimal.

The process of item development is summarized in Figure 2 below. Briefly, 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, (4) conducted think aloud interviews with eighth grade students to establish construct validity and identify smaller item flaws, (5) moderated detailed scoring guides for items based on pilot test administrations, (6) administered final items to 600+ eighth grade students, and (7) used IRT to analyze findings from large-scale administrations.

*Figure 2. Iterative item development process.*
PRELIMINARY FINDINGS

Item Response Theory: Partial Credit Model results.

After two annual cycles of the BEAR Assessment System, IRT analysis indicates that the current version of our learning progression for scientific argumentation has a degree of validity, particularly for the lower and intermediate levels of our model, and has provided insights into a better understanding of the higher levels where there is a good fit between the map and the item. This outcome is a product of an improvement in our knowledge of how to write argumentation items and improvement in our model of progression. As the focus of this paper is primarily on the building of our learning progression for argumentation and the methods for how it is operationalized, Osborne et al (2013, in prep) provides a more detailed analysis of the empirical data generated by our assessments.

IMPLICATIONS

The value of the work is twofold. First it offers a tested model of the trajectory of student expertise with argumentation. Perhaps more importantly, it offers a body of expertise of how to measure and assess student competency with argumentation in science. This is particularly important given the policy priority to assess what 21st century higher order thinking skills in the context of science in particular argumentation (National Research Council, 2011). As Pellegrino (2013) has argued:

Our greatest danger may be a rush to turn the NGSS into sets of assessment tasks for use on high-stakes state accountability tests before we have adequately engaged in research, development, and validation of the range of tasks and tools needed to get the job done properly.

If teachers are to develop students’ ability to engage in argumentation, we must provide the means to assess the competencies that we value and support the development of instructional environments that foster the use of scientific reasoning and argumentation. For in the absence of a well defined construct, items will be written that operationalize an ill-defined or nebulous construct or, even worse, the absence of items will lead to the omission or undervaluing of a practice which is central and core to understanding science.

REFERENCES


Cavagnetto, A. R. (2010). Argument to Foster Scientific Literacy A Review of


## Appendix A: Proposed Progress Map for Argumentation/Scientific Reasoning

<table>
<thead>
<tr>
<th>Level</th>
<th>Constructing Arguments (first person)</th>
<th>Critiquing Arguments (third person)</th>
<th>Description (The student can…)</th>
<th>Cognitive Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>N/A</td>
<td>N/A</td>
<td>No evidence of any facility with argumentation</td>
<td>![C</td>
</tr>
<tr>
<td>0a</td>
<td>Making a relevant claim</td>
<td>N/A</td>
<td>Explicitly state a <strong>claim</strong>. Does not have to be correct or substantiated, but must represent an arguable point.</td>
<td>![C]</td>
</tr>
<tr>
<td>0b</td>
<td>N/A</td>
<td>Identifying/stating the claim of/for someone else</td>
<td>Explicitly identify a <strong>claim</strong>. This could be in the form of a position or hypothesis.</td>
<td>![C]</td>
</tr>
<tr>
<td>0c</td>
<td>Stating evidence supporting your claim</td>
<td>Identifying/stating evidence supporting the claim of/for someone else</td>
<td>Explicitly identify a piece of <strong>evidence</strong> that can be used to <em>implicitly</em> support an argument.</td>
<td>![C - W - E]</td>
</tr>
<tr>
<td>1a</td>
<td>Stating a warrant for your argument</td>
<td>Identifying/stating a warrant for the argument of/for someone else</td>
<td>Explicitly identify a <strong>warrant</strong> that <em>implicitly</em> links evidence to a claim.</td>
<td>![C - W - E]</td>
</tr>
<tr>
<td>1b</td>
<td>Constructing an explicit argument</td>
<td>Identifying/constructing an explicit argument of/for someone else</td>
<td>Explicitly identify a <strong>claim</strong>, a piece of <strong>evidence</strong>, and a <strong>warrant</strong> that links that evidence to a claim.</td>
<td>![C - W - E]</td>
</tr>
<tr>
<td>1c</td>
<td>Constructing a rebuttal to another argument</td>
<td>Identifying/constructing a rebuttal of/for someone else</td>
<td>Explicitly identify a <strong>claim</strong>, a piece of <strong>evidence</strong>, and a <strong>warrant</strong> that links that evidence to a claim <em>in such a way that is superior to another argument</em>.</td>
<td>![C - W - E]</td>
</tr>
<tr>
<td>2a</td>
<td>Constructing a one-sided comparative argument</td>
<td>Identifying/constructing a one-sided comparative argument of/for someone else</td>
<td>Evaluate two competing arguments and provide explicit reasons for why one argument is superior.</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>---------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>2b</td>
<td>Making a two-sided comparative argument</td>
<td>Identifying/constructing a two-sided comparative argument of/for someone else</td>
<td>Evaluate two competing arguments and provide explicit reasons not only for why one argument is superior but also for why the other argument is inferior.</td>
<td></td>
</tr>
<tr>
<td>2c</td>
<td>Stating relative significance of several pieces of evidence to your claim</td>
<td>Identifying/stating relative significance of several pieces of evidence to the claim of/for someone else</td>
<td>Evaluate the various degrees of importance multiple pieces of evidence may have to a specific claim.</td>
<td></td>
</tr>
<tr>
<td>2d</td>
<td>Constructing an explicit counter claim</td>
<td>Identifying/constructing an explicit counter claim of/for someone else</td>
<td>Explicitly identify an alternative explanatory hypothesis and provide explicit justification for why it is superior to other competing arguments</td>
<td></td>
</tr>
</tbody>
</table>