Student Assessments for Reading and Writing Scientific Arguments

Amanda M. Knight[1], Katherine L. McNeill[1], Seth Corrigan[2], and Jacqueline Barber[2]

Boston College[1]

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

contact info:

Amanda M. Knight

Lynch School of Education, Boston College

140 Commonwealth Avenue, Chestnut Hill, MA 02467

Knightam@bc.edu

Reference as:

Knight, A. M., McNeill, K. L., Corrigan, S., & Barber, J. (2013, April). *Student assessments for reading and writing scientific arguments.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Abstract

The focus on scientific argumentation within the Next Generation Science Standards and Common Core Standards Initiatives reflects the acceptance of an expanded and more authentic perspective of competence. While comprehensive, effective, and scalable classroom tools and assessments for argumentation are needed to support these new initiatives, the field still lacks valid and reliable instruments to measure such competency. In this paper, we present a model for developing science assessments targeting this key scientific practice, including both construct maps and corresponding sample assessment items. We tend to the structural characteristics of arguments within two modalities—reading and writing.

Student Assessments for Reading and Writing Scientific Arguments

We are at an exciting moment in science education—one in which we have come to recognize an expanded and more authentic view of competence in which students are knowledgeable in reading and writing complex informational text as well as engaging in written and oral argumentation using the rules of evidence and reasoning that are respected in scientific discourse (Pearson, Moje, & Greenleaf, 2010). This perspective is the result of 20 years of research in which the significance of evidence-based arguments has been established not only as a central practice of scientists (Duggan & Gott, 2002) and medical practitioners (Aikenhead, 2005), but also of learning science (Duschl, Schweingruber, & Shouse, 2007; Osborne, 2010). In turn, argumentation has been recently incorporated into national standards for literacy (Common Core State Standards Initiative, 2010a), math (Common Core State Standards Initiative, 2010b) and science (Achieve, Inc., 2013) education. Therefore, there is a need to incorporate argumentation into numerous standards-based assessments; however the field lacks such valid and reliable instruments (Osborne, 2010). Consequently, we are working to address this gap. In this paper, we present a new vision of assessment items that seek to measure competency in reading and writing scientific arguments.

Arguments in Science and Literacy

Arguments for Explanations

By including science practices, such as evidence-based explanations and arguments, within the recently released Next Generation Science Standards (Achieve, Inc., 2013), the scientific education community is promoting an expanded view of science knowledge. No longer is science knowledge limited to the currently accepted explanatory accounts of

phenomena; rather, it also includes the practices that are used to establish, extend, and refine (NRC, 2012, p.26) that knowledge through conflict and argument (Latour, 1987). As such, arguments about evidence-based explanations support the development of both conceptual and epistemic knowledge (Osborne, Erduran, & Simon, 2004). By epistemic knowledge, we mean the values within the scientific community such as the preference for data as a form of justification (Sandoval & Cam, 2011) as well as the acceptable types of investigation questions and methods for collecting data (Sandoval & Reiser, 2004) that impact the persuasiveness of said arguments. While scientists argue about things other than explanatory accounts, such as deign solutions, methods, and models, our work focuses on arguments about evidence-based scientific explanations.

Yet, argumentation, the process of constructing and critiquing arguments about scientific claims, evidence, and alternative explanations (Driver, Newton, & Osborne, 2000; NRC, 2012), remains an uncommon classroom practice (D. Kuhn, 1993; Newton, Driver & Osborne, 1999) that can be challenging for students (Osborne et al., 2004). Specifically, research suggests that even when students know that they should justify their claims, they often have trouble doing so because they tend to not know how to critique the quality of scientific evidence (McNeill & Krajcik, 2007). As persuasion in science is dependent on the quality of scientific evidence (Berland & McNeill, 2012), students who tend not to or do not know to critique the evidence likely construct less persuasive arguments. As such, we need high quality instruments that will measure students' ability to critique the quality of scientific evidence in order to better support students.

Components of Arguments

Similar to other research (Erduran, Simon, & Osborne, 2004; Jiménez -Aleixandre, Rodriguez, & Duschl, 2000), we utilize a structural definition of argument that builds on the components from Toulmin's (1958) argument pattern. The components include: 1) Claim–an answer to the question, 2) Evidence–data (measurements or observations) and/or patterns, trends, and/or inferences from the data that support the claim, 3) Reasoning–a justification that uses scientific ideas to explain how or why the evidence supports the claim, and 4) Rebuttal–a justification for how or why an alternative explanation is incorrect (McNeill & Krajcik, 2012). While research suggests that arguments with more components are more sophisticated (Berland & McNeill, 2012; Clark & Sampson, 2008; Kuhn, 1991; Osborne et al., 2004; Schwarz, Neuman, Gil & Ilya, 2003; Voss & Means, 1991), we are examining how the quality of one of these components—evidence—impacts the overall sophistication of students' arguments. Moreover, we examine how students' abilities to critique arguments based on the quality of evidence is similar and different depending on the modality (reading, writing, and speaking).

Arguments Across Modalities

The modalities of reading, writing, and talking provide equally important, but different opportunities to engage in evidence-based arguments about explanations. Take, for instance, what it means to identify a claim within each modality. When reading one must locate the claim within the text to understand the meaning and purpose of an argument. In comparison, when writing one must make inferences about the data prior to being able to construct the claim. Lastly, when talking one presents a justified claim in response to a question, or listens and responds to the claim someone else is making. The latter of which could require the former.

While these skills are related, there are significant differences. For instance, whereas reading is primarily receptive, writing is primarily productive. Moreover, talking, which requires both listening and speaking, crosses both actions. Specifically, listening requires the reception of language, and speaking is a form of producing language. Therefore, it could be important to understand and discriminate between these in order to develop a sophisticated understanding of argumentation.

The scientific argumentation research has mostly focused on how best to support students in constructing written (e.g. McNeill, 2011; McNeill et al., 2006; Sampson et al., 2010; Sandoval & Millwood, 2005) and spoken arguments (e.g. Berland & Reiser, 2011; Osborne et al. 2004; Jiménez –Aleixandre et al., 2000; Sampson et al., 2010; Varelas, Pappas, Kane, & Arsenault, 2008). Much less research has examined students' abilities to critique scientific arguments when reading (e.g. Phillips & Norris, 1999; Norris & Phillips, 1994; Ratcliffe, 1999). While our project is designing assessments that measure students' abilities to critique arguments within the modalities of talking, reading, and writing, this paper focuses only on the latter two modalities.

Scientific Argumentation Assessments

Instructional Context

The goal of this project is to develop scientific argumentation assessments that will be embedded within a middle school curriculum unit that is currently under development through collaboration between The Learning Design Group at the Lawrence Hall of Science and Amplify Learning (www.amplify.com). The curriculum emphasizes a multimodal approach to learning science (Pearson et al., 2010) as well as the scientific practices identified in the Next Generation Science Standards (Achieve, 2013), including an emphasis on scientific explanations and

arguments. Therefore, the curriculum emphasizes constructing and critiquing arguments across the modalities of reading, writing and talking. Moreover, the curriculum will be delivered on a tablet computer (e.g. iPad), and, as such, the argumentation assessments will also be developed on a technology platform. We do, however, plan to also provide paper and pencil versions that can be implemented apart from the curriculum. The implementation of the argumentation assessments within this curriculum also dictated the content domains of the assessments. Specifically, we are developing life as well as earth science versions of the argumentation assessments. However, we maintain a goal to try to limit the content knowledge as much as possible from interfering with the assessment of students' argumentation abilities. While it is not possible to completely detach content from argument, we have built in scaffolds that try to minimize this interaction. This will be discussed in more depth when the development of the items is presented. The argumentation assessments presented in this paper are limited to the Earth Science version.

Assessment Design

We are designing the reading and writing assessments using the BEAR Assessment System (BAS) (Wilson, 2005; 2009) in conjunction with elements of evidence-centered design (Mislevy, Almond & Lukas, 2004). The BAS is comprised of iterative steps that include four building blocks: 1) Construct maps, 2) The item design, 3) The outcome space, and 4) The measurement model. Development begins with the design of construct maps—theoretical models of cognition that extend from high to low ability and illustrate qualitatively distinct groups of respondents and responses to items (Wilson, 2005). The second stage of development—item design—shows how the theoretical construct can be measured. While this

step focuses on the item stem, the third stage of development—the outcome space—addresses how students answer the question. For multiple-choice questions, the outcome space is often focused on developing strong answer choices based on theory. In comparison, the outcome space for constructed response items addresses the qualitative classification of student responses using rubrics or scoring guides. It is within the item development and outcome space that we also used elements of evidence-centered design (Mislevy et al., 2004) to clarify item specifications and ensure they result in valid inferences about student abilities. Specifically, we developed a design pattern for every level within each construct map. The design pattern defines the knowledge, skills, and abilities that would be evidence of a student's ability at each level of the construct, characteristic features of the task(s) that are designed to measure each level as well as potential observations of students' responses to constructed response items at each applicable level. The final stage of development is the measurement model, which relates individual scores and assessment items to the original construct map. We plan to cycle through these steps a minimum of three times, during which the construct maps, items, and outcome spaces will be revised based on evidence that either supports or weakens validity claims. A descriptive Item Response Modeling (IRM) approach will be used to examine the theorized ordering of the construct levels (DeBoeck and Wilson, 2004). However, as this is a conceptual paper, this paper will focus on the development of the constructs as well as their corresponding maps, items and outcome spaces.

Scientific Argument Constructs

There are numerous characteristics of arguments that could be measured. Each characteristic represents a construct that could be mapped into qualitatively distinct levels. The

constructs that we are developing within our project for the reading and writing modalities are presented in Table 1. This table represents a comprehensive list of what we have decided to develop based on a combination of available resources and what we believe will be most beneficial in supporting students argumentation skills; it does not represent a comprehensive list of all the possible scientific argument constructs. The constructs include: 1) forms of justification, 2) relevant-supporting evidence, 3) sufficiency of evidence, 4) multiple claims, and 5) text structure. We will next provide a very brief summary of each of these constructs.

First, the forms of justification construct assesses students' preference for the type of support used to justify the claim, such as empirical evidence (Berland & McNeill, 2012), science ideas (Osborne et al., 2004), appeals to authority, plausible mechanisms, and/or prior experiences (Sandoval & Cam, 2012). Second, the relevant-supporting evidence construct assesses whether students critique evidence (data and/or patterns, trends, and/or inferences from the data that support the claim) based on whether it fits with the claim (relevant) and exemplifies the relationship stated in the claim (support). Third, the sufficiency of evidence construct tracks whether students use all of the relevant-supporting evidence for a multivariate phenomenon. Fourth, the multiple claims construct distinguishes between students who attend to the main claim and those who recognize, consider, or critique an alternative claim. Fifth, the text structure constructs assesses students' ability to follow the organization of the argument.

While we are measuring some of the constructs in both reading and writing (e.g. forms of justification and relevant supporting evidence), others could be measured in both modalities, but we have not made them a priority within our work (e.g. sufficiency of evidence and multiple claims). Moreover, in some cases a particular modality makes affordances that are not possible within other modalities. Specifically, we were only able to develop items that address how

students' critique the organization of an argument within the reading modality (e.g. text structure). While the constructs that are unique to a modality are interesting and noteworthy, in this paper we have chosen to tease apart the similarities and difference that result when the same construct—relevant-supporting evidence—is measured across multiple modalities (e.g. reading and writing).

Table 1.

Construct maps under development in the writing and reading modalities.

Construct Maps	Reading	Writing
Forms of Justification	✔	✔
Relevant-Supporting Evidence	✔	✔
Sufficiency of Evidence	~	✔
Multiple Claims	~	✔
Text Structure	✔	✗

✔under development; ~ chosen not to develop; ✗ not appropriate for the modality

Relevant-Supporting Evidence

Relevancy and support impact the quality of scientific evidence, and, therefore, the quality of the argument as a whole (NRC, 2012). We define relevant evidence as data that addresses (or fits with) the claim. Relevant data has the *potential* to be of high quality if it is also supportive of the claim. Therefore, supporting evidence can be defined as evidence that exemplifies the relationship established in the claim. For instance, if a claim were based on a trend in the data (e.g. earthquake are more destructive when their focus is closer to the Earth's surface), supporting evidence would include data that exemplifies that trend (e.g. Earthquake's A and B were shallow and had a higher destructive power).

Data that is irrelevant or relevant-contradictory tend to weaken the overall argument. Irrelevant evidence, which is neither supportive nor contradictory, weakens an argument by introducing tangential ideas. The same can be said when relevant-contradictory, which supports a different or alternative claim, is used as support within a main argument because it does not exemplify what the claim is purporting (e.g. it is not supporting).

Reading Construct Map. Our reading relevant-supporting evidence construct map (see Table 2) was developed from the literature as well as from the expertise of our team. In regards to the literature, there is very little research that has examined students' abilities to read scientific arguments. Specifically, the findings from one research team suggest that high school students' struggle to identify evidence when reading science news articles (Phillips &Norris, 1999; Norris & Phillips, 1994). This suggests that locating evidence within a text could be problematic for students, and, therefore, these studies informed the lower border of our construct map. Specifically, students whose ability is at level 1 are able to locate evidence when reading a scientific argument, whereas the ability of students who are not able to locate the evidence is level 0.

Informing the upper border of our construct map, another research study found that middle school students tended to be able to provide relevant critiques of extrapolations made from the evidence they read in a scientific news article (Ratcliffe, 1999). This is supported by other argumentation literature that posits critique as a difficult skill (Osborne et al., 2004). Therefore, it is not until levels 3 and 4 that students are able to critique the relevant-supporting evidence. Whereas the students whose ability is on target with level 4 can critique relevant-supporting evidence across two arguments, students whose ability is on target with level 3 can only critique relevant-supporting evidence within a single argument. At the highest level,

however, students will critique the evidence in two arguments with the same claim, but with different evidence. These arguments are similar to the extrapolations that the middle school students in Ratcliffe's (1999) study tended to be able to critique.

However, our experience also tells us that there seems to be a wide gap between the difficulty of locating relevant-supporting evidence (level 1) and critiquing relevant-supporting evidence (levels 3 and 4). We postulate that this in-between step would be to select new relevant-supporting evidence when provided with multiple options. Therefore, the students who are able to do this are on target with level 2.

Table 2.

Relevant-supporting evidence construct map for the reading modality

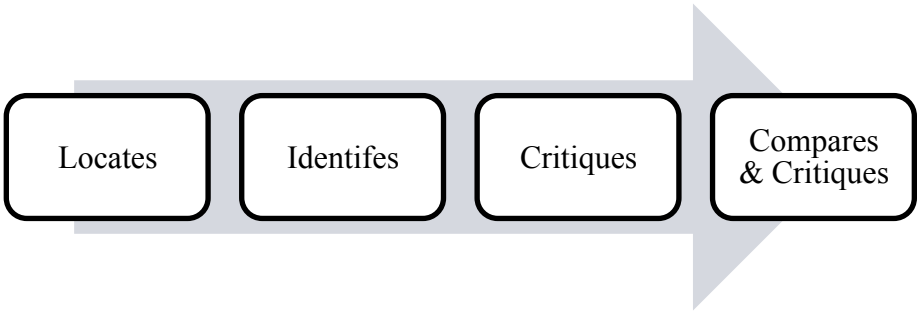
	Levels	Description	Items
High	**Compares & Critiques** Spivey & King, 1989 (Synthesis) Osborne et al., 2004	Student critiques the quality of the evidence in terms of relevancy and support when comparing two arguments.	Item 4
	Critiques Osborne et al., 2004	Student critiques evidence based on both relevance and support.	Item 3
	Identifies Kintsch & Van Dijk, 1978 (Interpret) Spivey & King, 1989 (Categorization)	Student identifies relevant-supporting evidence.	Item 2
	Locates Kintsch & Van Dijk, 1978 (Locate)	Student locates the evidence of an argument.	Item 1
Low	**Does not locate**	Student does not locate the evidence of an argument.	Item 1

The verbs used across the reading construct map were chosen purposefully (e.g. locates,

identifies, critiques, and compares & critiques). The distinction between locate and identify is based on Kintsch and Van Dijk's (1978) model of text comprehension, which positions locate and recall as being of lower level than interpret. Our application of identifies parallels Kintsch and Van Dijk's application of interpret as well as Spivey and King's (1989) application of categorization within their read to write model. Moreover, Spivey and King posit that organization is more difficult than categorization, and synthesis is more difficult than organization. While we do not have a level that maps onto organization, Spivey and King's application of synthesis, is similar to our highest level of critique in that it requires the reading across two texts. The application of critique at the upper border is also supported by research within the scientific argumentation literature. Specifically, Osborne and his colleague (2004) incorporated critique of the structural components of an argument at the highest levels of their framework. Taken together, the research supports our decision to use locate at the lower border and critique at the upper border, with identification falling in the middle.

Figure 1.

Increasing sophistication of verbs on the reading construct maps.



Reading Items and Outcome Spaces. All of the reading items are dichotomous, which target the ability associated with a single attribute level. For instance, if a student gets the

answer correct, then their ability corresponds to the difficulty level of the item. However, if the student answers the question incorrectly, then we only know that there ability is less than the difficulty level of the item. Consequently, we developed items that correspond to the locates, identifies, critiques, and compares and critiques levels of the reading relevant-supporting evidence construct map. We did not need to develop an item at the does not locate level as students whose ability is at this level will incorrectly answer the locate item.

Each of the four items are bundled together into an item-set (see Appendix A). Moreover, they all require use of the same information that is presented at the beginning of the item-set. Specifically, the item-set begins with a causal wonderment question that addresses a univariate phenomenon, followed by a dataset as well as a sample student's argument. In the example item-set provided (see Appendix A), the wonderment question is: Why do some volcanoes have larger eruptions than others? The empirical data consists of names of five volcanoes, quantitative data for the outcome variable (amount of magma), and qualitative data for the dependent variable (amount of time between eruptions). We chose to use both qualitative and quantitative data because some researchers have found that students tend not to recognize qualitative data as data that can be counted as evidence (Sandoval & Millwood, 2005).

Every item-set follows the same characteristics: five data entries, data for the outcome variable is quantitative, and data for the dependent variable is qualitative. The sample student argument consists of 3 sentences: 1) Claim, 2) Relevant-supporting evidence, and 3) Reasoning. To identify the sentences, we used the notation (S1), (S2), and (S3) to indicate sentence 1, 2, and 3 respectively. To make this notation as clear as possible, we included a legend prior to the sample student's argument. This notation is aligned with how other assessment systems are developing items (e.g. PARCC). We will next discuss how each item uses the same information

differently.

Item 1: Locates. The purpose of the locate item is to determine whether a student can find the sentence that includes evidence when they read the sample student's argument. The answer choices were purposefully chosen: Sentence that contains the claim, sentence that contains the evidence, sentences that contain the claim and evidence, and none. Due to the complexity of the underlying mechanisms and the grade levels for which we are developing these assessments (middle school), the reasoning was often simplistic and easily confused with the claim or evidence. Therefore, we purposefully decided to not include it as an answer-choice to this question.

Item 2: Identifies. In this item type the student is required to decide which answer choice includes relevant-supporting evidence. In addition to relevant-supporting evidence and relevant-contradictory data, the answer choices included two irrelevant data options. The relevant evidence included data on the time between volcanic eruptions and the associated amount of magma. The relevant evidence was supporting if it denoted an inverse relationship: the trend that less magma is released when the volcano erupts more often (or vice versa). In comparison, the relevant evidence was contradictory if it denoted a direct relationship: the trend that less magma is released when the volcano erupts less often (or vice versa). Each of the irrelevant data options only discusses one of the pertinent variables and often expounds on some tangential information. For instance, the first irrelevant option focused only on the frequency of eruptions (21 volcanoes have erupted in the last 20 years in the US), and provided tangential information around the number of active volcanoes in the US. In comparison, the second irrelevant option focused only on the amount of magma released. Neither of these is relevant

because to be relevant the evidence would need to include data or a pattern of the data for both the time between volcanic eruptions and the amount of magma released.

Item 3: Critiques. For this item type students are asked to critique a new piece of evidence. In the item provided, the new piece of information is relevant-supporting; however, this varied by item-set and could also be relevant-contradictory or irrelevant. In the outcome space the student is required to make a judgment about the quality of the evidence based on its relevance and support. We used "addresses the question" as a proxy for relevancy. Because the new evidence is relevant-supporting in this example, the correct answer choice is "excellent because it provides support for the claim and addresses the question". The other options choices included not supporting and irrelevant, not supporting and relevant, and supporting and irrelevant. The answer choices always remained the same. The correct answer depends on whether the new evidence is relevant-supporting, relevant-contradictory, or irrelevant. If the new evidence is relevant-contradictory, the correct answer choice is supporting and irrelevant. If the new evidence is irrelevant, then the correct answer choice is not supporting and irrelevant.

Item 4: Compares & Critiques. In order to compare and critique the relevant-supporting evidence within two arguments, the fourth item introduces a new sample student argument. Students are now required to compare the evidence used within the new argument to evidence used within the first sample student's argument. The two arguments always use different evidence: relevant-supporting, relevant-contradictory, or irrelevant. In the level 4 item provided the first argument uses relevant-supporting evidence and the second argument uses irrelevant evidence. The outcome space then requires that students make a judgment about which sample student's evidence is stronger. Because the first argument uses relevant-supporting evidence that student's evidence is stronger (Margaret). Similar to the level 3 item, the outcome space also

requires the student to critique the relevancy and support, however with the level 4 item the student must critique the evidence in both arguments. Therefore, to answer the question correctly the student must know 1) Margaret's evidence provides support for the claim and addresses the question (relevant-supporting), 2) Winston's evidence does not address the question and does not support the claim (irrelevant), and 3) Relevant-supporting evidence is better than irrelevant evidence (Margaret's evidence is stronger).

Writing Construct Map. Similar to the reading relevant-supporting evidence construct map, the writing relevant-supporting evidence construct map (see Table 3) was informed by the research literature. At the lowest level of this construct map, the student does not provide any evidence (level 0). Students, whose ability is at level 1, include some irrelevant and non-supporting data in addition to relevant-supporting evidence. The research on students' abilities to use evidence when writing scientific arguments suggests that students usually try to use data as evidence (Sandoval & Millwood, 2005), but routinely use irrelevant evidence (L. Kuhn & Reiser, 2005; McNeill & Krajcik, 2007; Sandoval, 2003). By using both qualitative and quantitative data, we can assess whether students value one type of data of another. When students are able to limit their evidence to that which is only relevant to and supportive of the claim, then their ability is at level 3. To be able to limit the evidence, suggests that the student had to first critique the evidence. Critique, therefore, occurs at the upper border, which is consistent with scientific argumentation research (Osborne et al., 2004).

Table 3.

Relevant-supporting evidence construct map for the writing modality

	Level	Description	Item
High	Relevant	Student critiques and limits all of the empirical evidence to that which supports and is relevant to the claim.	Item 1
↕	Mixture	Student provides a mixture of relevant and/or supporting empirical evidence as well as irrelevant and/or non-supporting data.	
Low	None	Student does not provide any empirical evidence (observations or measurements that support the claim).	

Writing Items and Outcome Spaces. Whereas each reading relevant-supporting evidence item corresponds to a single level within one particular construct map, a student's response to one writing item is used to assess the his/her ability level across every writing construct map (e.g. forms of justification, relevant-supporting evidence, sufficiency of evidence, and multiple claims). As such, we had to pay close attention to the item format in order to be sure that all constructs could be scored within each item.

A sample-writing item (item 6) is presented in Appendix B. Similar to the reading items, the writing items begin with a wonderment statement that introduces students and their question. For instance, in item 6 Joe and Bob wonder why some earthquakes have more destructive power than others. An authority statement, a mechanism statement, and empirical data follow the question. Each of these will next be discussed in further detail.

An authority statement in which an expert, such as a scientist, provides information related to the question follows the wonderment statement. The authority statement is always three sentences. While the first sentence introduces a scientist and their area of study (e.g. Dr. Schmidt visited the students' class, and explained that she studies that affect islands), the second introduces a science fact that could be used to answer the question (e.g. They learn that right now

she is studying the Haiti 2010 earthquake, and that the city called Port-au-Prince is where the destructive power was the greatest). The third sentence then introduces a picture of the event the scientist is studying (e.g The picture below shows the damage after the 2010 earthquake in Port-au-Prince, Haiti). The goal of the authority statement is to provide another form of justification that students could use to answer the question or a second claim that could be rebutted. The forms of justification construct dictates the need for this section. If a student only uses information from this section then we can infer that he/she values authority over the other forms of justification (e.g. mechanism or data).

Next, we provide a mechanistic statement that support the students' scientific reasoning and selection of relevant variables in the data table, which aligns with our goal to measure students' argumentation abilities as removed from content as possible. The mechanism is always three sentences. The first sentence introduces the relevant variables (e.g. Bob learns that some earthquakes happen deeper inside the Earth than others, and that earthquakes can happen in different ground materials), and the remaining two sentences explain the underlying mechanism (how the depth of the earthquake's focus and the hardness of the ground affect how the earthquake's waves travel trough the Earth). It is important to note that the mechanism alone cannot be used alone to answer the question because it does not provide the direction of the relationship between the variables (e.g. direct or inverse relationship). Rather, this relationship always requires an inference from the data table (e.g. "when waves begin closer to the Earths surface" is a proxy for the depth variable). Similar to the authority statement, if students only draw from this section, we can infer that it is a form of justification that he/she most values.

Lastly, we provide an empirical data chart. The data chart includes five entries, the outcome variable, an irrelevant variable, and either one or two dependent variables depending on

whether it is a univariate or multivariate phenomenon. Item 6 is a multivariate question (e.g. both depth and hardness of ground affect the destructive power of the earthquake). As was also the case for the reading items, the outcome variable (e.g. destructive power) for the writing items is quantitative. This is also true for the irrelevant variable. We chose to represent the dependent variables (e.g. depth and hardness of ground) as qualitative in nature. Again, this is because some researchers have found that students tend not to recognize qualitative data as data that can be counted as evidence (Sandoval & Millwood, 2005). By using both qualitative and quantitative data, we can assess whether students value one type of data of another. The selection of the irrelevant variable was limited only to something that the students would recognize as being measurable, and the data does follow a trend from high to low or vice versa.

The entire item stem (wonderment statement, authority statement, mechanism statement, and empirical data) is organized onto one sheet of paper in the testing booklet. On the opposite faced page the question is provided again, and the remainder of the space is provided for the students' response. The student's response is scored using a specific rubric for each construct map. Table 4 presents the relevant-supporting evidence specific rubric for item 6. The main purpose of this rubric is to provide examples of what counts as relevant (e.g. "Both Earthquake A and B happened at a shallow depth", "Both Earthquake A and B happened in soft ground material", "Earthquake A had a larger destructive power") and irrelevant data (an average crust temperature of 51 ^{0}F, an average crust temperature of 77 ^{0}F).

Table 4.

Specific rubric to measure relevant-supporting evidence for writing item 6.

Levels	Description	Potential Observation
Relevant	Student limits all of the empirical evidence to that which supports and is relevant to the claim.	*Provide empirical evidence of depth and/or hardness of ground material and/or power* • "Both earthquake A and B happened at a shallow depth" • "Both earthquake A and B happened in soft ground material." • "Earthquake A had a larger destructive power."
Mixture	Student provides a mixture of relevant and/or supporting empirical evidence as well as irrelevant and/or non-supporting data.	*Provide relevant empirical evidence of depth and/or hardness of ground material (see above) AND irrelevant Crust temperature data.* • "Earthquake A happened at a shallow depth, with an average crust temperature of 51^0 F, in soft ground material, and had a destructive power of 12." • "Earthquake E happened at a deep depth, with an average crust temperature of 77^0 F, in very hard ground material, and had destructive power of 6."
None	Student does not provide any empirical evidence.	*No empirical evidence [data (observations or measurements) or patterns, trends and/or inferences from the data].*

Conclusion

The question still outstanding is: *How will teachers use this information?* We have explained how we developed the item design and outcome space for items that map onto each level of the construct. The progression of difficulty of the items maps onto the progression of difficulty found within the construct map. Therefore, if a student answers items 1 and 2 correctly on the reading assessment, but misses items 3 and 4, then his/her ability would be consistent with level 2 (identifies relevant-supporting evidence). Knowing this, a teacher could target his/her instruction specifically to helping this student learn how to critique relevant-supporting evidence. Having a suite of student assessment items that differentiate students' abilities to justify claims with evidence and reasoning may be an important tool for teachers to plan future instruction (Gotwals & Songer, 2010). Consequently, in the next phase of this project we will be tying appropriate instructional strategies to every level of each construct map.

This, however, begs another question: *Should we be supporting students differently in different modalities?* We have made the argument that the progression of students' abilities within a single construct is mapped very differently for different modalities. Otherwise stated, the construct is the same, however the construct maps are different. If there are differences (as we are suggesting), it is reasonable that the supports should vary depending on the modality. Moreover, a students' ability within the same construct may not be equivalent in two modalities. Specifically, several studies have found that there are differences between the quality of students' oral and written arguments (Berland & McNeill, 2012; Knight & McNeill, in review; Sampson et al., 2010). As such, a teacher would need to support arguments within the modalities differently.

References

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

Aikenhead, G. S. (2004). Science-based occupations and the science curriculum: Concepts of evidence. *Science Education, 89,* 242–275.

Bell, B. (2007). Classroom assessment of student learning. In S. Abell, & N. Lederman (Eds.), *Handbook of Research on Science Education* (pp. 965-1006). Mahwah, NJ: Lawrence Erlbaum Associates.

Berland, L. K. & McNeill, K. L. (2012). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, 94(5), 765-793.

Berland, L. K. & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education,* 93(1), 26-55.

Clark, D., & Sampson, V. (2006). The quality of argumentation supported by personally-seeded discussions. In T. Koschmann, T.W. Chan, & D. Suthers (Eds.), Computer supported collaborative learning 2005. Mahwah, NJ: Erlbaum.

Clark, D., & Sampson, V. (2008). Assessing dialogic argumentation in online environments to relate structure, grounds, and conceptual quality. Journal of Research on Science Teaching, 45(3), 293-321.

Common Core State Standards Initiative. (2010a). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http:// www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Common Core State Standards Initiative. (2010b). *Common core state standards for Mathematics*. Retrieved from http:// www.corestandards.org/assets/CCSSI_MATH%20Standards.pdf

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

De Boeck, P. & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades k-8.* Washington D.C.: National Academy Press.

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.

Duggan, S. & Gott, R. (2002). What sort of science education do we really need? *International Journal of Science Education*, 24, 661 – 679.

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*, 259-281.

Jiménez -Aleixandre, M. P., & Erduran, S. (2008). Designing argumentation learning environments. In S. Erduran & M. Jimenez-Aleixandre (Eds.), Argumentation in science education: Perspectives from classroom-based research (pp. 91-116). New York: Springer.

Jiménez -Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). 'Doing the lesson' or 'doing science': Argument in high school genetics. Science Education, 84(3), 287- 312.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension. *Psychological Review,* 85(5), 363-394.
Knight, A. M., & McNeill, K. L. (in review). Comparing students' verbal and written scientific arguments. *Science Education*.
Kuhn, D. (1991). The skills of argument. Cambridge, England: Cambridge University Press.
Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education, 77* (3), 319-337.
Kuhn, L., & Reiser, B. (2005). Students constructing and defending evidence-based scientific explanations. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, TX.
McNeill, K. L. (2011). Elementary students' views of explanation, argumentation and evidence and abilities to construct arguments over the school year. *Journal of Research in Science Teaching*, 48(7), 793-823.
McNeill, K. L., Corrigan, S., Barber, J., Goss, M. & Knight, A. M. (2012, March). *Designing student assessments for understanding, constructing and critiquing arguments in science.* Poster presented at the annual meeting of the National Association for Research in Science Teaching, Indianapolis, IN.
McNeill, K. L., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett & P. Shah (Eds.), *Thinking with data: The proceedings of the 33rd Carnegie symposium on cognition.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
McNeill, K. L. & Krajcik, J. (2012). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence and reasoning framework for talk and writing.* New York, NY: Pearson Allyn & Bacon.
McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, 15(2), 153-191.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to Evidence-Centered Design. CSE Technical Report 632, The National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE). LA, CA: University of California, Los Angeles.
Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education*, 21(5), 553-576.
Norris, S. P., & Phillips, L. M. (1994). Interpreting pragmatic meaning when reading popular reports of science. Journal of Research in Science Teaching, 31(9), 947-967.
Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328,* 463-466.
Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching,* 41(10), 994-1020.
Pearson, P. D., Moje E. B., & Greenleaf, C. (2010) Literacy and Science: Each in the Service of the Other. Science, 328, 459-463
Phillips, L. M., & Norris, S. P. (1999). Interpreting popular reports of science: What happens when the reader's world meets the world on paper? International Journal of Science Education, 21(3), 317-327.
Ratcliffe, M. (1999). Evaluation of Abilities in Interpreting Media Reports of Scientific Research. International Journal of Science Education, 21(10), 1085-1099.

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

Sampson, V. Grooms, J. & Walker, J. P. (2010). Argument-driven inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education,* 95(2), 217-157.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences*, 12, 5-51.

Sandoval, W. A., & Cam, A. (2011). Elementary children's judgments of the epistemic status of sources of justification. *Science Education*, 95(3), 383-408.

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

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, 88(3), 345-372.

Schwarz, B. B., Neuman, Y., Gil, J., & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activity. *Journal of the Learning Sciences*, 12(2), 219-256.

Spivey, N. N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Resarch Quarterly*, 24(1), 7-26.

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

Varelas, M., Pappas, C. C., Kane, J. M., & Arsenault, A. (2008). Urban primary-grade children think and talk science: Curricular and instructional practices that nurture participation and argumentation. *Science Education*, 92, 65-95.

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, 1, 337-350.

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

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716-730.

Appendix A: Relevant-supporting evidence items on the reading assessment

Mr. Pete asks his students to write an argument about the following question: **Why do some volcanoes have larger eruptions than others?** Margaret used the data table below to write her argument.

Name of Volcano	**Amount of Magma** (km^3)	**Amount of Time Between Eruptions**
Erebus	0.0001 — 0.001	Constantly erupting
Soufrière Hills	0.001 — 0.01	Few months
Mount Vesuvlus	0.1 — 1.0	Decade
Mount Pinatubo	1 — 10	Century
Mazama	10 — 100	Millennium

Margaret's Argument:

(S1) = Sentence 1 *(S2)* = Sentence 2 *(S3)* = Sentence 3

(S1) I think that volcanoes that release a lot of magma do not erupt very often. (S2) The volcano called Mazama last erupted over a millennium ago and released more than 10 km^3 of magma. (S3) Volcanoes that release a lot of magma do not erupt very often because it takes more time to build up enough gas pressure needed for really big eruptions.

1. **In which statement does Margaret support her argument with evidence?**

   a. Sentence 1 (S1)

   b. Sentence 2 (S2)

   c. Sentence 1 (S1) and Sentence 2 (S2)

   d. None

2. **Margaret is thinking of adding more evidence to her argument. Which piece of evidence best supports her claim?**

    a. There are 169 active volcanoes in the United States and 21 have erupted in the last 20 years.

    b. The volcano called Unzen tends to erupt every few months, and releases about 0.01 km^3 of magma.

    c. The volcano called Kīlauea has been erupting since 1983 and has released about 3.5 km^3 of magma.

    d. The Guarapuava Volcano released 8,600 km^3 of magma, which is a large amount of magma.

3. **Denzel is also in Mr. Pete's class. He is thinking about adding a piece of evidence to his argument that volcanoes that release a lot of magma do not erupt very often.**

    The volcano called Tambora tends to erupt about once every millennium and releases about 100 km^3 of magma.

    This piece of evidence is

    a. poor because it does not provide support for the claim or address the question.

    b. fair because it addresses the question, but does not support the claim.

    c. fair because it supports the claim, but does not address the question.

    d. excellent because it provides support for the claim and addresses the question.

Winston is also in Mr. Pete's class. Mr. Pete asked Margaret and Winston to compare arguments to see who used stronger evidence.

Margaret's Argument:

> *I think that volcanoes that release a lot of magma do not erupt very often. The volcano called Mazama last erupted over a millennium ago and released more than 10 km^3 of magma. Volcanoes that release a lot of magma do not erupt very often because it takes more time to build up enough gas pressure needed for really big eruptions.*

Winston's Argument:

> *Volcanoes that release a lot of magma do not erupt very often. There are more than 1,500 volcanoes that have erupted around the world in the past 10,000 years. 150 of these volcanoes are located in the United States. A few of these have produced some of the largest and most dangerous eruptions anywhere in the world during this century.*

4. **Which student, Margaret or Winston, better supports his or her argument? Why?**

    a. Margaret's evidence is stronger. She provides support for the claim and addresses the question. Winston's evidence is weaker. He addresses the question, but does not support the claim.

    b. Margaret's evidence is stronger. She provides support for the claim and addresses the question. Winston's evidence is weaker. He does neither.

    c. Winston's evidence is stronger. He provides support for the claim and addresses the question. Margaret's evidence is weaker. She supports the claim, but does not address the question.

    d. Winston's evidence is stronger. He provides support for the claim and addresses the question. Margaret's evidence is weaker. She does neither.

Appendix B: Writing assessment item

Item 6.

Joe and Bob wonder why some earthquakes have more destructive power than others

Dr. Schmidt visited the students' class, and explained that she studies earthquakes that affect islands. They learn that right now she is studying the Haiti 2010 earthquake, and that the city called Port-au-Prince is where the destructive power was the greatest. The picture below shows the damage after the 2010 earthquake in Port-au-Prince, Haiti.

Bob learns that some earthquakes happen deeper inside the Earth than others, and that earthquakes can happen in different ground materials. He also learns that earthquakes travel through the Earth in waves, and the waves are largest where the earthquake happens**.** When the waves begin closer to the Earth's surface and when the waves can move through the Earth more easily, they have more destructive power.

Joe found the table below:

Earthquake	Destructive Power at Epicenter	Average Crust Temperature 1 mile Below Surface (^{0}F)	Depth	Hardness of Ground
A	12	51	Shallow	Soft
B	10	53	Shallow	Soft
C	8	59	Intermediate	Soft
D	7	65	Intermediate	Hard
E	6	77	Deep	Very hard

Item 6:

Joe and Bob need your help. Using the information Joe and Bob learned about earthquakes, write an argument that answers the question:

What causes some earthquakes to be more destructive than others?
How do you know?