Assessing Middle School Students' Abilities to Critique Scientific Evidence

Amanda M. Knight[1], Cecilia Brito Alves[2], Matthew A. Cannady[2], Katherine L. McNeill[1], &

P. David Pearson[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., Alves, C. B., Cannady, M. A., McNeill, K. L., & Pearson, P. D. (2014, April). *Assessing middle school students' abilities to critique scientific evidence.* Paper presented at the annual meeting of NARST, Pittsburg, PA.

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 focus on one specific construct, relevant-supporting evidence within the reading modality, and present empirical results from a pilot study supporting the assessment model.

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 critiquing relevant-supporting evidence while reading scientific arguments. Moreover, we present results from one pilot study that provide validity evidence for our assessment model.

Theoretical Framework

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. While it

is possible to argue about many different things within the field of science (e.g., models, design solutions, evidence, methods, explanatory accounts, etc.), our work focuses specifically on arguments about explanations. While arguments use evidence to *defend* how or why a phenomenon occurs, explanations *make sense* of why a phenomenon occurred (Berland & McNeill, 2010). For instance, an explanation might clarify why the biodiversity decreased, while an argument might focus on defending why one explanation for why the biodiversity decreased is the most appropriate explanation. A focus on arguments about explanations represents an expanded view of science knowledge that includes the currently accepted explanatory accounts of phenomena as well as 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 (e.g., core ideas within NGSS) 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.

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 arguments of lower quality. As such, the science education community needs valid and reliable instruments that will measure students' ability to critique the quality of scientific evidence in order to better support students with constructing and critiquing scientific arguments.

Components of Arguments

In our work we utilize the claim, evidence, reasoning, rebuttal (CERR) framework, which focuses on the structure of scientific arguments (McNeill & Krajcik, 2012). Specifically, we look for the presence and quality of specific components within the argument. The components include: 1) Claim–a statement that answers the question, 2) Evidence–data (measurements or observations) and/or patterns, trends, and/or inferences from the data that justify 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; D. Kuhn, 1991; Osborne et al., 2004; Schwarz, Neuman, Gil & Ilya, 2003; Voss & Means, 1991), we examine how the quality of one of these components—evidence—impacts the overall sophistication of students' arguments. Specifically, we examine how students' abilities to critique arguments based on the quality of evidence.

Reading Arguments

Our view of scientific literacy aligns with Norris & Phillips's (2003) fundamental sense

of science literacy, which they define as reading and writing when the content is science (Norris & Phillips, 2003). Scientific argumentation is a unique disciplinary literacy practice in that it can and should be applied across the modes of reading, writing, and talking. For instance, students should read and critique others' arguments as well as construct their own written or oral arguments. In this study, we examine more closely how students read and critique scientific arguments.

Reading is often cast as "the passive receipt of information" (Pearson et al., 2010, p. 460). Science educators, therefore, tend to view reading as a less desirable replacement to scientific inquiry (Pearson et al., 2010), which includes "asking questions, planning and conducting investigations, using appropriate tools and techniques to gather data, thinking critically and logically about relationships between evidence and explanations, constructing and analyzing alternative explanations, and communicating scientific arguments" (NRC, 1996, p.105). However, when reading is cast as an inquiry-driven practice, then it becomes a "processes of actively making meaning of science" that can be used to support scientific inquiry (Pearson et al., 2010, p. 460). From this critical perspective, reading becomes a tool to investigate phenomena in ways that help students learn how to use other scientists' methods and findings as a starting place for their own investigations (Cervetti, Pearson, Bravo, & Barber, 2006). Therefore, inquiry-based reading can promote scientific inquiry, including scientific argumentation.

While research has examined students' abilities to construct their own written (e.g. McNeill, 2011; McNeill, Lizotte, Krajcik, & Marx, 2006; Sampson, Grooms, &, 2010; Sandoval & Millwood, 2005) and oral arguments (e.g. Berland & Reiser, 2011; Osborne et al. 2004; Jiménez –Aleixandre, Rodriguez, & Duschl, 2000; Sampson et al., 2010; Varelas, Pappas, Kane,

& Arsenault, 2008), very little research has examined students' abilities to critique scientific arguments when reading (e.g. Phillips & Norris, 1999; Norris & Phillips, 1994; Ratcliffe, 1999). This study focuses specifically on students' abilities to read and critique the evidence in scientific arguments.

Critiquing the Quality of Evidence Based on Relevance and Support

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 science matter of 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 higher destructive power, whereas Earthquakes C and D were deep and had lower destructive powers).

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. Tangential ideas weaken an argument because they draw the reader away from the point of the argument. Thus, the argument becomes less persuasive. The same can be said when relevant-contradictory evidence, which supports an alternative claim on the same science topic, is used as support within a main argument because it does not exemplify the relationship the claim is purporting (e.g. it is not supporting).

The research has tended to focus on students' abilities to construct and critique scientific

evidence suggests that it can be challenging for students. Although students often try to use data to support their claims (Sandoval & Millwood, 2005), they routinely use irrelevant evidence (Kuhn & Reiser, 2005; McNeill & Krajcik, 2007; Sandoval, 2003). Moreover, students rarely interpret the meaning of evidence or explain why it counted as evidence (Sandoval & Millwood, 2005). Students also tend not to recognize observations as relevant qualitative evidence, nor did that typically reference lack of data as evidence that was relevant to discount claims (Sandoval & Millwood, 2005). It is important to find ways to assess students' abilities to construct and critique scientific evidence because of the significant role scientific evidence plays in establishing the quality of a scientific argument.

Reading Relevant-supporting Evidence Assessment

Assessing Argumentation

Assessments are used to measure what students know about a construct. A construct is the latent (unobservable) characteristic that is being measured, which for this study is students' abilities to critique the quality of relevant-supporting evidence when reading scientific arguments. Assessments are based on the process of gathering evidence about students' knowledge and abilities, and use this evidence to make inferences about what students know (Mislevy, Wilson, Ercikan, & Chudowsky, 2002). With the recent incorporation of argumentation as one of the eight essential scientific practices within the Next Generation Science Standards (Achieve, Inc., 3013), there is a need to incorporate argumentation into numerous standards-based assessments. However, the field does not currently have any valid and reliable instruments to measure scientific argumentation (Osborne, 2010).

While Berland and McNeill (2010) proposed a hypothetical learning progressions for oral

and written scientific argumentation, it has not been empirically validated. A learning progression is a sequence of more complex ways of thinking that develop over time (Smith, Wiser, Anderson, & Krajcik, 2006) and are designed to support conceptualization and development of assessments (Wilson, 2009). Similarly, our work initiated with the development of theoretical learning progressions that compared the expressive modalities (writing and speaking) to the receptive modality of reading (McNeill, Corrigan, Goss, & Knight, 2012). We moved, however, towards developing smaller slices of the learning progression, which are known as construct maps. Wilson (2009) explains "one straightforward way to see the relationship of construct map to learning progression is to see the learning progression as composed of a set of construct maps, each comprising a ''dimension'' of the learning progression" (p.723). Otherwise stated, the construct map is a piece of the larger learning progression.

Assessment Design

We are designing the reading and writing assessments using the BEAR Assessment System (BAS) (Wilson, 2005; 2009). 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. The final stage of development is the measurement model, which relates individual scores and assessment items to the original construct map. Specifically, we employ Masters' (1982) partial credit Rasch model, which will be further explicated in the data analysis for the pilot study. The results to two previous pilots were used to refine the construct maps, items and outcome spaces.

Reading Relevant-Supporting Evidence Construct Map. Our reading relevant-supporting evidence construct map (see Table 1) 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 critique the quality of evidence in scientific arguments they have read. 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 are below level 1.

Informing the upper border of our construct map, the argumentation literature posits critique as a difficult skill (e.g. Osborne et al., 2004). While the findings from one study suggest that some middle school students were able to critique extrapolations made from the evidence they read in a scientific news article (Ratcliffe, 1999), other studies suggest that critiquing evidence can be challenging. Specifically, students' tend to use irrelevant evidence (Kuhn &

Reiser, 2005; McNeill & Krajcik, 2007; Sandoval, 2003) and rarely interpret the meaning of evidence or explain why it counted as evidence (Sandoval & Millwood, 2005) when constructing scientific arguments. Therefore, it is not until levels 3 and 4 that students are able to critique the quality of relevant-supporting evidence. Whereas the students whose ability is on target with level 4 can compare the quality of two arguments based on critiques of the relevance and support of the evidence, students whose ability is on target with level 3 can only critique the quality of evidence based on relevance and support within a single argument.

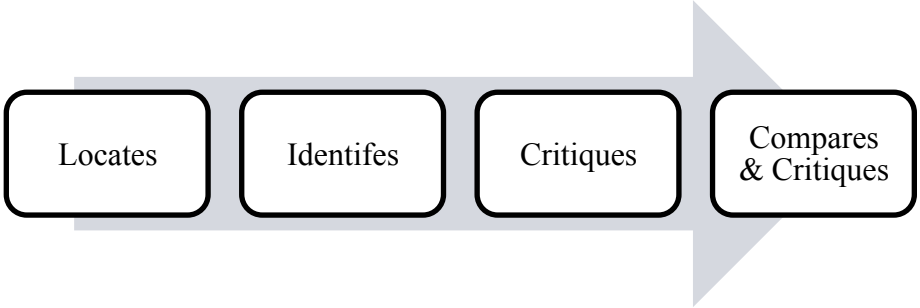
However, our experience also tells us that there seems to be a wide gap between the difficulty of locating relevant-supporting evidence (level 1) and critiquing relevant-supporting evidence (levels 3 and 4). We postulate that this in-between step would be to select new relevant-supporting evidence when provided with multiple options. This is supported by research that suggests that students tend not to recognize observations as relevant qualitative evidence or reference lack of data as evidence that was relevant to discount claims (Sandoval & Millwood, 2005). Therefore, students who are able to identify relevant-supporting evidence are on target with level 2.

Table 1.
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
Low ↓	**Locates** Kintsch & Van Dijk, 1978 (Locate)	Student locates the evidence of an argument.	Item 1

The verbs used within the reading construct map were purposefully chosen (e.g., locates, identifies, critiques, and compares & critiques) (see Figure 1). 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 reading across two texts.

Figure 1.
Increasing sophistication of verbs on the reading construct maps.



Reading Relevant-Supporting Evidence Items and Outcome Spaces. We developed items that correspond to the locates, identifies, critiques, and compares & critiques levels of the reading relevant-supporting evidence construct map. Each item, therefore, targets the ability associated with a single construct level. For instance, if a student gets the answer correct, then their ability corresponds to at least the construct level associated with that particular 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.

Each set of four items consists of three multiple-choice items and one constructed response item, and all the items use the same introductory information. We refer to set of four items as a testlet. Specifically, the testlet begins with a wonderment question that addresses a univariate phenomenon, followed by a dataset as well as a sample student's argument. In the example testlet provided (see Appendix A), the wonderment question is: What is related to the height of ash clouds that are released from volcanoes? The empirical data consists of names of five volcanoes, quantitative data for the outcome variable (explosive power of volcano), and qualitative data for the dependent variable (height of ash cloud). 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). We also provided an irrelevant variable (wind speed) that used quantitative data with no apparent pattern. We chose to include an irrelevant variable to help them to consider that not all data is evidence.

We developed four different testlets, each consisting of four items. While the topic of two of the testlets focuses on volcanoes, the other two testlets focus on earthquakes. Each testlet 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 introductory information in

different ways and for different purposes.

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. 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 a direct relationship: the trend that higher ash clouds correspond to more powerful eruptions (or vice versa). In comparison, the relevant evidence was contradictory if it denoted an inverse relationship: the trend that higher ash clouds correspond to less powerful eruptions (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 size of the ash cloud, and provided tangential information about stopping planes from flying in parts of Europe for six days. In comparison, the second irrelevant option focused only on explosive power of eruptions, and provided tangential information about the largest recorded eruption. 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 power of the eruption and the height of the ash cloud.

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 testlet 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. Because the new evidence is relevant-supporting in this example, the correct answer choice is "good because it supports Ben's claim". 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 (Ben). Similar to the level 3 item, the outcome space also requires the student to critique the arguments based on the relevancy and support of the evidence, 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) Ben's evidence provides support for the claim and addresses the question (relevant-supporting), 2) Anna's evidence does not address relevant science and does not support the relationship in the claim (irrelevant), and 3) Relevant-supporting evidence is better than irrelevant evidence (Ben's evidence is stronger). Students' responses to this item were scored on four levels (0 to 3) using a scoring guide (see Appendix B).

Pilot Study

Research Questions

The goal of our pilot study was to determine whether the ordering of our theorized construct levels was appropriate, and whether we needed to make any refinements to the levels or items. Consequently, we ask:

1. What are students' abilities to critique the quality of relevant-supporting evidence (RSE) when reading scientific arguments?
2. What aspects of the reading RSE assessment need refinement?

Methods

Participants. This study uses data from a national pilot in which teacher participation was solicited through various listservs. Eight teachers agreed to participate in the study, and were compensated with a $50 Amazon gift card for their time and effort in attaining student assent, parent consent, and administering the test. In total, 679 students took the assessment. Their locations are summarized in Table 2. Moreover, each student answered items within 2 testlets. Table 3 summarizes the number of students who took each item.

Table 2.
Summary of participants.

Partner Districts	Number of Students
Flagstaff, AZ	91
Redding, CA	61
Cataula, GA	100
Rockford, IL	87
Greensboro, NC	76
Wells River, VT	19
Oak Harbor, WA	139
London, England	106
TOTAL	679

Data Collection. The reading relevant-supporting evidence construct has four levels (e.g., locates, identifies, critiques, compares & critiques), and items were developed to correspond to each level. Each set of four items consists of three multiple-choice items and one constructed response item, and all the items use the same introductory information. We refer to set of four items as a testlet. This pilot employed four different testlets, each consisting of four items[1]. While the topic of two of the testlets focuses on volcanoes, the other two testlets focus on earthquakes. Each student in this pilot was assigned two testlets (I=8), and each testlet was completed on a separate day. To assign students two testlets, the testlets were combined into 12 different forms based on the possible combinations (see Table 3). A particular forms was assigned to each teachers' class, with the goal of having a minimum of 50 students complete each form, which is a minimum of 200 students per item. The assessment was presented and answers collected via an online survey program (i.e., Qualtrix), and the students used either tablets or computers to answer the items. Table 4 summarizes the number of students who answered each item. Missing data was handled as a pairwise deletion. While this allowed us to use more of the data, there are different numbers of students who answered each item.

[1] We had two items that did not fit the model, RSE_01_L3 and RSE_04_L3, and, as such, they were removed from the analysis.

Table 3.

Number of students who took each form of the reading RSE assessment.

Form	Testlet		Teacher	Class	Number of Students	
	Day 1	Day 2			Total per class	Total per form
1	Earthquake 1	Earthquake 2	1	A	26	44
			5	A	19	
			9	A	25	
2	Earthquake 1	Volcano 1	1	B	21	65
			5	B	19	
			9	B	25	
3	Earthquake 1	Volcano 2	1	C	14	69
			6	A	30	
			9	C	25	
4	Earthquake 2	Earthquake 1	2	A	27	58
			6	B	31	
5	Earthquake 2	Volcano 1	2	B	23	53
			6	C	30	
6	Earthquake 2	Volcano 2	2	C	30	47
			7	A	17	
7	Volcano 1	Earthquake 1	2	D	30	47
			7	B	17	
8	Volcano 1	Earthquake 2	2	C	29	46
			7	C	17	
9	Volcano 1	Volcano 2	4	A	21	63
			7	D	18	
			3	A	24	
10	Volcano 2	Earthquake 1	4	B	21	65
			7	E	18	
			3	B	26	
11	Volcano 2	Earthquake 2	4	C	23	68
			7	F	19	
			3	C	26	
12	Volcano 2	Volcano 1	4	D	22	47
			9	A	25	
			9	A	25	

Table 4.
Number of students that answered each item

Item	Testlet			
	Earthquake 1	Earthquake 2	Volcano 1	Volcano 2
1: Locate	268	278	247	203
2: Identify	268	278	247	204
3: Critique	269	277	247	204
4: Compare & Critique	290	296	252	201

Data Analysis. The constructed response items were scored using the scoring guide (see Appendix B) by two independent raters with 25% overlap and 85% reliability. Before conducting Rasch analyses, we first checked the assumption of unidimensionality by examining the factor loadings on a full information factor analysis, explained variance, scree plot, and biserial correlation. The use of more than one criterion is inline with how most researchers determine the dimensionality of a test (Costello & Osborne, 2005). In regards to the Rasch analyses, we used ConQuest (Wu, Adams, & Wilson, 1997) to apply the Masters' (1982) partial credit model to both the multiple choice and coded scores for the constructed response items, which expresses the probability of a randomly selected participant, X_i, with ability level θ to obtain a score of x on item i and category k of difficulty δ_{jk} correctly, and is defined by the following function:

$$P_{ik}(\theta) = P(X_i = x|\theta)\frac{\exp\sum_{j=0}^{k}(\theta-\delta_{jk})}{\sum_{j=0}^{m-1}\exp\sum_{j=0}^{i}(\theta-\delta_{jk})}.$$

The partial credit model (Masters, 1982) is appropriate when analyzing polytomous items with different numbers of response options. Polytomous means that the students' abilities can be categorized in more than two levels, which occurred within the constructed response items in which the students' responses were categorized into one of four levels (0 to 3). Furthermore, we

have a different number of response options because we employed both multiple choice and constructed response items. The multiple-choice items have two response options (e.g., 0=wrong; 1=right), whereas there are four possible response options (e.g., scores corresponding to the four levels of the rubric: 0, 1, 2, 3) for the constructed response items. Similarly, Wu, Adams, and Wilson (1997) apply the partial credit model for analyzing data from the Third International Mathematics and Science Study (TIMMS). The TIMMS tests consist of 158 test items, which is a mixture of multiple choice, short answer, and extended response items. In this study they "specified the partial credit model because it will deal with the mixture of dichotomous and polytomous items" (pg. 82). More information about this issue can be found on Chapter 7 from the Conquest Manual (Wu, Adams, & Wilson, 1997).

Assumptions

The assumption set forth in Rasch modeling require that the data fit the Rasch model as well as that the test items and examinees confirm to the model. This means that the students' abilities and item difficulties are measured on the same scale, which affords comparisons across. However, there are other assumptions that should be addressed prior to running Rasch analyses: item discrimination and unidimentionality.

Item Discrimination. One assumption within Rasch modeling is that all items are equally discriminating, which can be assessed through the point biserial correlation (pbis). The pbis is the correlation between student performance on a particular item in relationship to the total test score. Pbis is employed as an index to examine item discrimination, that is, how well an item can discriminate between the high and low performing students. According to Lietz (1995), a high correlation occurs when there is "a strong link between the item and the scale" (p.

166). The lower the discrimination index, the weaker the pattern of linkage between the item and the scale. Theoretically, the pbis correlation ranges from -1 to +1, but in practice this range is restricted to -0.20 to 0.75 (du Toit, 2003). Moreover, values above 0.20 are considered satisfactory. As shown on Table 5, which presents the point biserial values, all items have pbis values above the minimum cut off of 0.20. The smallest pbis was found for a Level 4 item RSE02_L4 (pbis=0.28) and the highest for a Level 2 item RSE01_L2 (pbis=0.66). In general, the test possesses high discriminative power with average pbis equal 0.48.

Table 5.
Point biserial

Item	Point Biserial
RSE02_L1	0.48
RSE02_L2	0.35
RSE02_L3	0.51
RSE02_L4	0.28
RSE06_L1	0.45
RSE06_L2	0.65
RSE06_L3	0.34
RSE06_L4	0.50
RSE04_L1	0.46
RSE04_L2	0.53
RSE04_L4	0.39
RSE01_L1	0.51
RSE01_L2	0.66
RSE01_L4	0.57

Dimensionality. Another assumption of Rasch modeling is that items are unidimensional. Simply stated, this means that the items should only measure one variable. Most researchers seldom use a single criterion in determining the dimensionality of a test (Costello & Osborne, 2005). A number of methods for identifying the number of factors have been proposed (see for example, Embretson and Reise, 2000). In an effort to check the

assumptions of unidimensionality of the reading assessment, four indices were used: factor loadings on a full information factor analysis, explained variance, and scree plot.

Factor Loadings. These provide information about how strongly items are related to the latent construct. Items with stronger relationship to the construct are deemed more reliable indicators of that construct (Edwards & Wirth, 2009), meaning student scores on these items are seen as indicators of overall student ability on the intended construct being measured. To determine whether an item is a component of a specific factor, a cutoff value of 0.3 (Lambert & Durand, 1975) is recommended as an acceptable minimum value for the coefficients. The loadings presented on Table 6 were computed using the Full Information approach. The main advantage of this approach is that it uses all available information in the estimation procedure (i.e., analyzes all item response patterns instead of analyzing each item separately). This approach was implemented using the Testfact software, which is based on the full-information item factor analysis proposed by Bock, Gibbons, and Muraki (1988). The average factor loading for the one-factor solution was 0.43, and only three items had had loadings lower than 0.30. Factor loadings lower than 0.3 suggest that more than 90% of the variance in an observed variable is explained by factors other than the construct to which the variable should be theoretically related[2].

[2] Assuming unidimensionality, as these factor loadings are from a 1-factor model run.

Table 6.
Factor loadings.

Item	Factor Loading
RSE02_L1	0.36
RSE02_L2	0.20
RSE02_L3	0.36
RSE02_L4	0.03
RSE06_L1	0.37
RSE06_L2	0.81
RSE06_L3	0.23
RSE06_L4	0.43
RSE04_L1	0.32
RSE04_L2	0.56
RSE04_L4	0.41
RSE01_L1	0.32
RSE01_L2	0.96
RSE01_L4	0.69

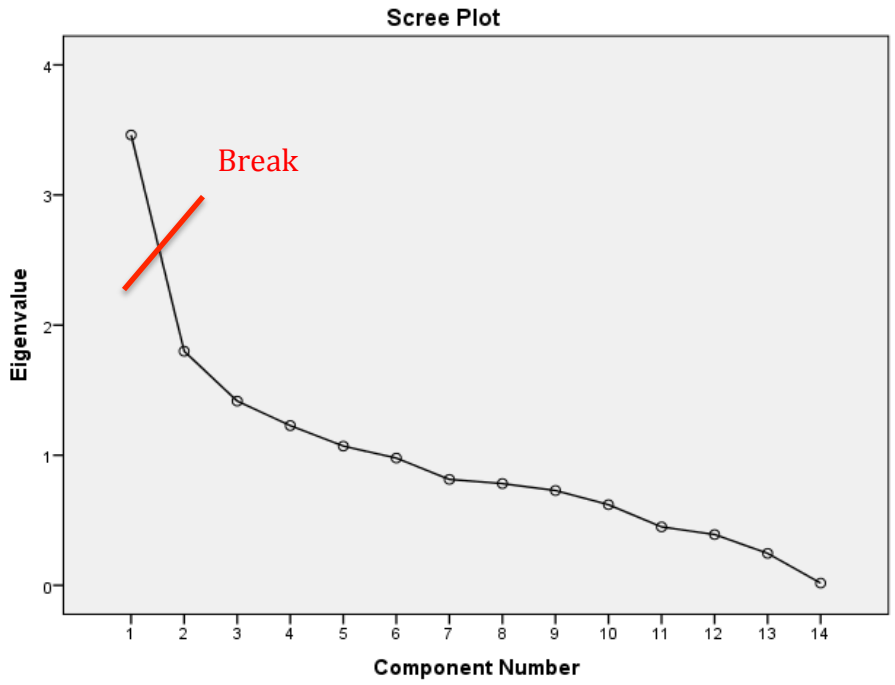
Explained Variance. The percentage of variance criterion is an approach based on achieving a specified cumulative percentage of total variance extracted by successive factors. The purpose is to ensure practical significance for the derived factors by ensuring that they explain at least a specified amount of variance. The optimal number of factors can be defined as the minimum number of factors that accounts for the maximum possible variance. Table 7, which presents the total variance explained, suggests that the reading test could possibly be explained by five factors with eigenvalues larger than 1.0. However, accepting five factors may harm the interpretability of the factors as they do not exhibit a conceptual meaning. As advocated by Daultrey (1976), parsimony is a very desirable feature in component analysis. Consequently, we argue that the one factor solution is the most meaningful solution.

Table 7.
Total variance explained.

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total
1	3.461	24.724	24.724	3.036
2	1.799	12.853	37.576	2.477
3	1.416	10.116	47.693	1.922
4	1.228	8.771	56.463	1.553
5	1.070	7.644	64.108	1.614
6	0.978	6.986	71.094	
7	0.814	5.814	76.909	
8	0.782	5.586	82.495	
9	0.729	5.206	87.701	
10	0.620	4.426	92.127	
11	0.449	3.208	95.335	
12	0.391	2.790	98.124	
13	0.246	1.755	99.879	
14	0.017	0.121	100.000	

Scree plot. Another common method employed to detect factors is the use of a scree plot. The position of a break or discontinuity in the pattern of eigenvalues can be suggestive of the number of factors to keep (Cattell, 1966; Tabachnick and Fidell, 2007). As shown in Figure 2, the first component has an eigenvalue approximately 3.5. In addition, the elbow, which indicates the sharpest break in the size of the eigenvalues, is found for between components 1 and 2. This evidence may suggest that the reading relevant-supporting evidence construct map could best be described using a single factor. While this method has been criticized because there may be no clear way to identify the "breaks" or "discontinuity" and because of the subjective method of determining where the break falls (Kim & Mueller, 1978), our data indicates one very apparent break.

Figure 2.
Scree plot



Taking into account all the evidence regarding the dimensionality, we argue that a one factor solution makes the most practical significance and we can affirm our instrument is a valid measure of the intended construct, students' abilities to read relevant-supporting evidence. Because these results suggest unidimensionality, the use of a Rasch analysis is appropriate.

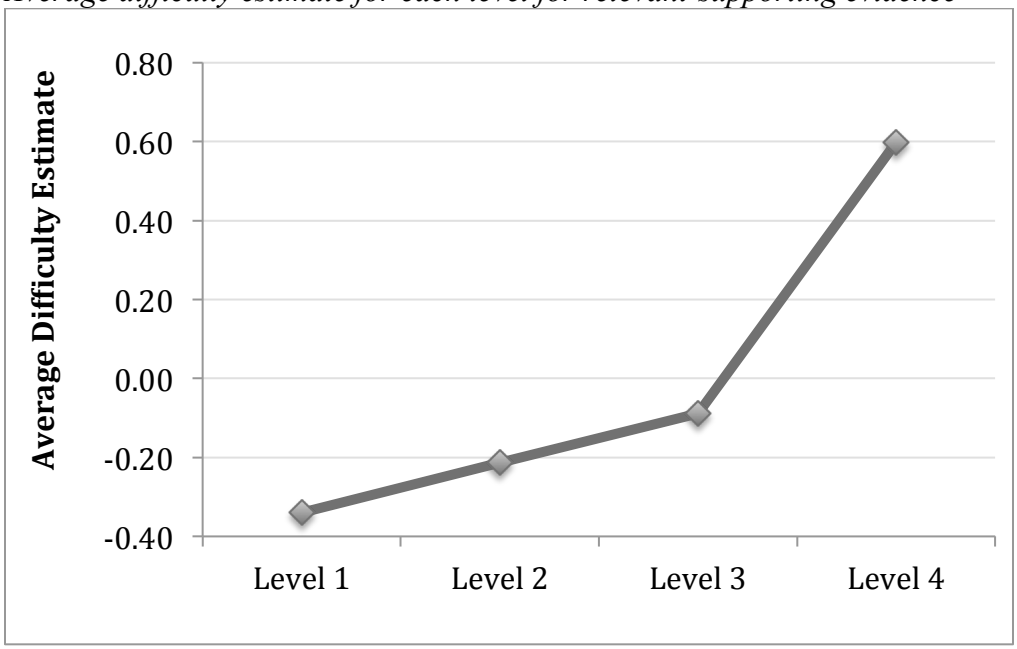
Internal Consistency

The Cronbach's alpha, which is a measure of internal consistency, was 0.68. This is just below the suggested minimum of 0.70. A high level of internal consistency is anticipated when all the items making up an instrument are intended to measure a single, unidimensional construct (ATS, 2007). Therefore, a value of 0.68 strengthens our argument that this is a single construct.

Results

Average Difficulty Estimates. The average difficulty estimates for each level of the reading relevant-supporting evidence construct map are presented in Figure 3. From Figure 3, we see that the ordering of the difficulty of the four levels is consistent with the hypothesized model on Table 1, where levels should increase in difficult in the following order: Level 1 → Level 2 → Level 3 → Level 4. As expected, on this construct map, the average difficulty of items making up Level 1 is lower than Level 2, 3 and 4. The first two levels can be considered easy, L3 moderate, and L4 hard. While the average difficulty estimates do show the overall difficulty trend across the construct map, they do not provide further information in regards to what is going on within each construct level. As such, we will next examine the Wright map.

Figure 3.
Average difficulty estimate for each level for relevant-supporting evidence



Wright Maps. Rasch IRT Wright maps measure students' abilities and item difficulties on the same scale, which affords comparisons between student ability and item difficulty. Figure

4 provides a Wright map of the students' ability and item difficulty parameters. The left-hand side of this figure represents the distribution of the measured ability of the students. The most proficient students register at the top of the graph and the least proficient at the bottom. Each "x" represents 3.4 students. Moreover, the right-hand side of this figure represents the distribution of the measured difficulty of the items. Again, the items of highest difficulty register at the top of the graph, whereas the items of lowest difficulty are at the bottom.

The first thing that should be noted in Figure 4 is that the student ability follows an approximate normal curve. However, we would like to see the normal distribution of students' abilities within the range of the difficulty of the items. Because there are a number of students whose ability level is below that of the item of lowest difficulty, this is considered a difficult test. Moreover, this also suggests that the students are not well separated, which is reflected in the person separation reliability of 0.470. The person (and item) separation reliabilities (see Table 9) have a maximum of 1.0 and a minimum of 0.0, and values close to 1.0 indicate that the person parameters are well separated and covering a range of the latent variable. As separation is a measure of the standard error of the estimates to the magnitude of the error in the estimates (Wright & Stone, 1979), we are striving to get a large standard deviation and small standard errors. A low person separation reliability of 0.470, therefore, indicates that the error term is larger than desired. This implies that our study sample was too narrow in ability range to adequately separate using our four-item scale. Either additional response options or a greater range in person ability are needed to increase person separation. First, in terms of the number of response options, the error term increases when students respond to fewer items (Linacre, 2012). Consequently, both the length of the test as well as missing data can inflate the error term. As each student only answered two testlets (8 items) and we excluded two items from the model, it

is logical that the number of response options is negatively impacting the person separation reliability. Second, the ability range of the students impacts the person separation reliability, with the person separation reliability increasing when the sample includes students that have both extreme high and low abilities (Linacre, 2012). As previously mentioned, this assessment was difficult for students in our sample as evidenced by the large number of student abilities falling below the item of lowest difficulty. However, the students in our sample were not required to have knowledge around argumentation or evidence, and the piloting of the formative assessment was not associated with a curricular unit. Consequently, it is logical that we have few students with high scores, and that this is also impacting the person separation reliability.

Table 9. Person and item separation reliability

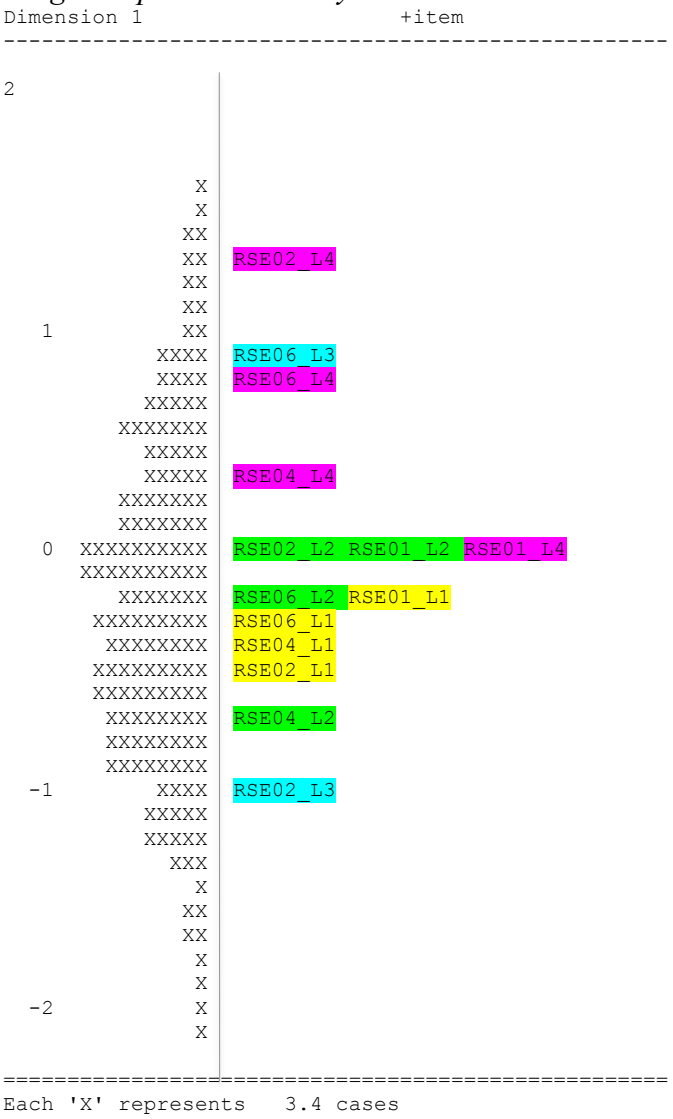
Person Separation	Item Separation
0.471	0.989

In comparison, the item separation reliability was 0.989, which suggests that our items are well separated and the sample is large enough to locate the items on the latent variable (Wright & Stone, 1979). This is also reflected on the Wright map in that there are minimal clumps and gaps within the range of item difficulties. As is usually the case, our item separation reliability is stronger than the person separation reliability, which is related to the amount of data associated with each. Specifically, there was more response data for the items (many students answered each item; N>200) than for the students (fewer items answered by each student; I=8).

Lastly, the construct levels with Figure 4 are color coded by item difficulty. According to our theory this should increase Level 1 → Level 2 → Level 3 → Level 4, however it is generally following the pattern Level 1 → Level 2 → Level 4 and it is not clear where Level 3 falls.

Figure 4.

Wright map color coded by construct level



While the construct levels are indicated after the item number, this view of the variable map does not make clear whether the items are ordering properly within each testlet. Therefore, we reorganized the variable map (see Figure 5) to separate the items by construct level and color-coded the testlets within each construct level. Specifically, the left column corresponds to level 1 (locate), the second column corresponds to level 2 (identify), the third column corresponds to level 3 (critique), and the rightmost column corresponds to level 4 (compare & critique). In

general, this organization provides the opportunity to see whether the attribute levels as a whole are increasing in difficulty from 1 to 4 as well as the ordering and spread of the items within each construct level. Moreover, in Figure 4, level 4 was presented as an average for each item. However, level 4 is a constructed response item consisting of 3 levels. In Figure 4, we include the three levels within this compare and critique level.

In looking at Figure 5, we do see a general increase from construct level 1 to level 4. Moreover, within level 4, we see that the rubric levels increase in difficulty from level 1 to level 3. However, there are two issues of note: similar difficulties between levels 1 and 2 as well as the large spread of item difficulties within level 3.

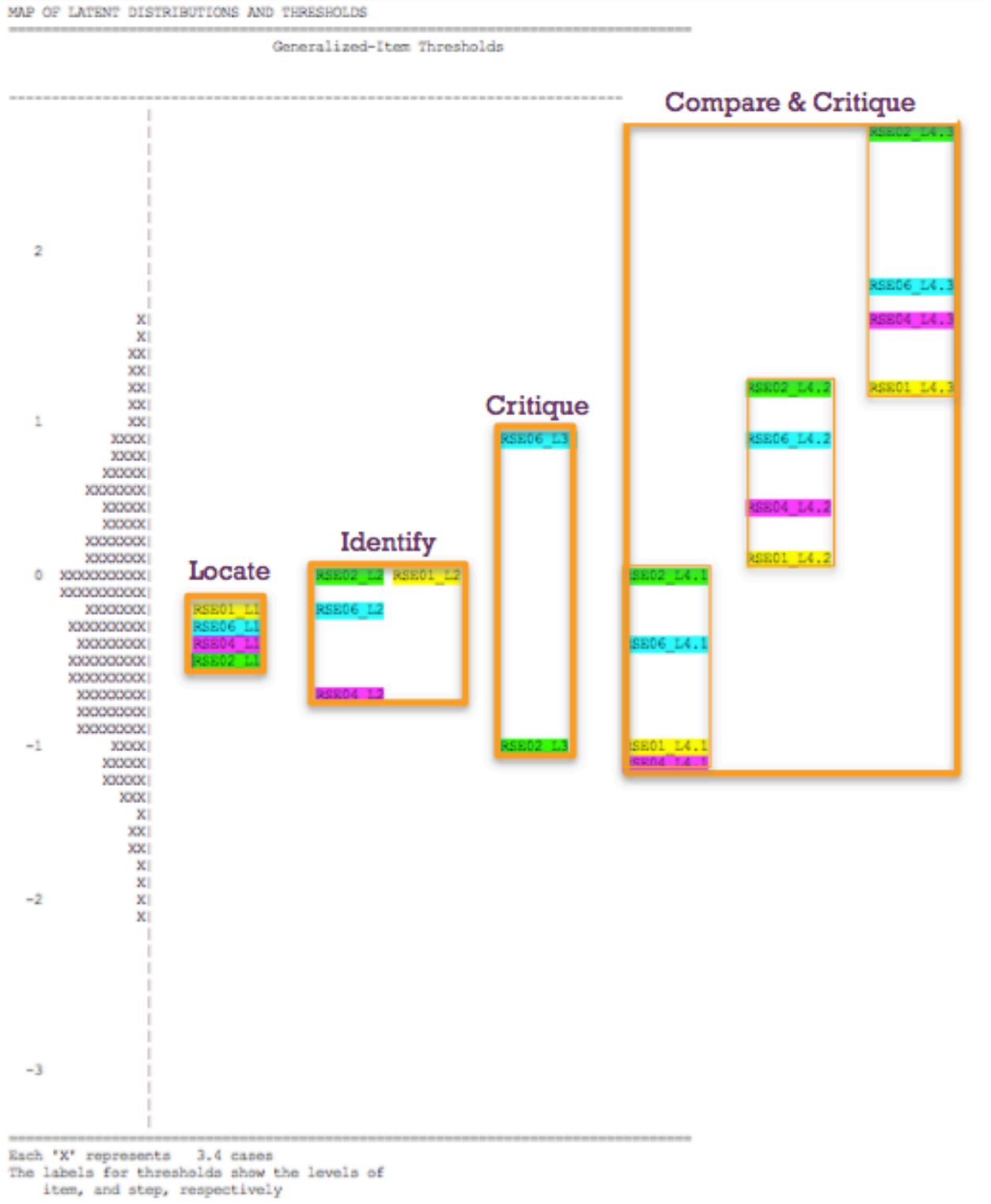
First, the difficulty of the items within levels 1 and 2 are very similar. While we theorized that locate would be easier than identify, the empirical results suggest otherwise, and we are not gaining any additional information from level 1 items above and beyond level 2 items. It is possible that this is because the locate items may not be measuring relevance and support of evidence. Cognitive interviews in which students think aloud and explain the reasons behind their answers could be done to further explore whether this is true. If we were to edit level 1, it perhaps should focus on identifying relevant evidence. This should be less difficult than identifying relevant-supporting evidence (level 2).

Second, there is a large spread in the difficulty of the items within level 3. Moreover, the two items that were removed from this model were both from this level (RSE_01_L3 and RSE_04_L3). This suggests that for this pilot, the level 3 items are not functioning as theorized. This, perhaps, could be due to the formatting of the item. Specifically, we tried to ask a critique question in a multiple-choice format, whereas it may be more appropriate to measure this through a constructed response format. That being said, in two previous pilots, we did not have

an issue with the spread within the level 3 multiple-choice items. Consequently, the spread could be an artifact of the pilot design. As opposed to students being randomly assigned to one of the twelve test forms, classes were randomly assigned to one of twelve test forms. As such, higher-level classes may have taken RSE_06_L3 and lower-level classes RSE_02_L3. If this is true, then the spread could be due to a class effect. In future pilots, this sampling issue will be rectified. Keeping these things in mind, it is important to also examine whether any of these items (level 1 or level 3) or other items have unexpected responses (misfit). This will next be discussed.

Figure 5.

Wright map arranged by construct level and color coded by testlet



Fit Statistics

It is also important to examine evidence in support of item fit: the weighted (infit) and unweighted (outfit) MNSQ. While each is a measure of misfit (i.e., unexpected responses), the outfit is sensitive to unexpected observations by examinees on items that are either very easy or very difficult for examinees whereas the infit statistic is sensitive to unexpected responses by examinees to items targeted at their ability level (i.e., are close to their ability level). In this analysis, we will consider both the infit and outfit statistics, and we assess the MNSQ significance for both infit and outfit statistics through ranges provided within the literature as well as through the t value.

The expected item MNSQ is a value of 1.0. Values less than 0.75 are considered significantly overfitting (Bond and Fox, 2007), which suggests the items are redundant with other items and are not providing unique information about the construct. In comparison, item MNSQ values greater than 1.0 indicate items are underfitting, which is a lack of construct homogeneity in relationship to the other items in a scale (Green, 1996). However, more specific cutoffs based on sample size (Smith, Schumacker, & Bush, 1998) are often applied. Specifically, items are considered underfitting when the unweighted and weighted mean square values are greater than 1.3 for samples less than 500, 1.2 for samples between 500 and 1000, and 1.1 for samples larger than 1000. As each item in this study was answered by 199 to 269 students, we apply the 1.3 standard. Last, t-values can be used to indicate significance of both infit and outfit MNSQ statistics. Specifically, t values above 2.0 indicate that the MNSQ infit and outfit statistics are significant.

From Table 10, which summarizes the item statistics, we see that no item shows a misfit, where the value of the weighted (infit) and unweighted (outfit) fit MNSQ are greater than the

critical value of 1.3 (underfit) or smaller than 0.75 (overfit). However, in looking at the t statistics, we observe three items (RSE02_L4, RSE06_L2, RSE01_L2) that have values greater that 2.0, which indicates at least one unexpected response (misfit). While item RSE02_L4 has a significant t-value for outfit (i.e., at least one unexpected response to an item that, based on the item difficulty and student ability, the student should have clearly answered correctly or incorrectly), items RSE06_L2 and RSE01_L2 have significant t-values for infit (i.e., at least one unexpected response to an item whose difficulty was near the student's ability level). The negative t-values for RSE06_L2 and RSE01_L2 merely indicate that more students than expected answered these items correctly. As the t-statistic is greatly affected by sample size (i.e., one unexpected response in a small sample can result in significance), the criterion is often changed to greater than 3 for smaller samples. Because the MNSQ ranges did not indicate underfitting or overfitting, and because we have a relatively small sample size as well as t-statistics lower than 3 for all items, we are less concerned about significance of items RSE02_L4, RSE06_L2, and RSE01_L2.

Table 10.
Difficulty estimates and Fit statistics for the relevant-supporting evidence items

Item	Estimate	Error	Outfil	T	Infit	T
RSE02_L1	-0.53	0.07	1.01	0.10	1.01	0.10
RSE02_L2	0.04	0.07	1.09	1.00	1.08	1.80
RSE02_L3	-0.99	0.07	0.99	-0.10	0.99	-0.20
RSE02_L4	1.29	0.06	1.27	2.90	1.17	1.80
RSE06_L1	-0.27	0.07	1.00	0.10	1.01	0.20
RSE06_L2	-0.18	0.07	0.89	-1.20	0.91	-2.30
RSE06_L3	0.82	0.07	1.10	1.20	1.07	1.00
RSE06_L4	0.74	0.06	1.00	0.10	0.99	-0.10
RSE04_L1	-0.36	0.07	1.03	0.40	1.03	0.70
RSE04_L2	-0.74	0.07	0.91	-1.00	0.93	-1.50
RSE04_L4	0.29	0.06	0.98	-0.20	0.99	-0.10
RSE01_L1	-0.20	0.07	1.00	0.10	1.00	-0.10
RSE01_L2	0.03	0.07	0.84	-1.70	0.87	-2.70
RSE01_L4	0.06	0.25	0.91	-0.90	0.92	-0.90

Summary

In general the overall Rasch solution was as expected. The difficulty of the construct levels increased from level 1 to level 4. The instrument will identify placement of students' abilities on the construct map, but refinements are still needed. First, Level 1, which currently focuses on locating relevant-supporting evidence, should be modified or removed as we were not gaining any additional information from level 1 that level 2 already provided. Second, we need to further investigate what happened within level 3. Specifically, was the spread due to item format (i.e., MC) or class effect? Third, the item separation and reliability call for a longer test with more items per attribute. While this would increase the sensitivity of the instrument in distinguishing between high and low level performers, we do not believe this is necessary for the low-stakes formative assessment in which a shorter length may be more useful. That being said, if level 1 is in fact not part of this construct, the internal consistency would likely increase if these items were removed.

Discussion

For students, the practice of argumentation (Osborne et al., 2004), which includes the ability to critique relevant-supporting evidence (Kuhn & Reiser, 2005; McNeill & Krajcik, 2007; Sandoval, 2003; Sandoval & Millwood, 2005), can be challenging. Likewise, this practice is also difficult to teach (Osborne et al., 2004), and yet, the field is by no means saturated with tools (e.g., curriculum, assessments) to support teachers with these challenges (Osborne, 2010). Aimed at filling this gap, in this paper, we presented one assessment model based on a progression for critiquing relevant-supporting evidence while reading scientific arguments, and supported this model with evidence from one pilot study. We believe this to be a meaningful progression that posits identification as important step towards critique. It is our hope that teachers can use this progression to inform their instruction as well as other researchers as a starting place in further conceptualization of this construct.

Meaningful Progression

Overall, the results from this study suggest that considering the relevance and support of evidence is a challenging skill for students, which aligns with previous research (e.g., Kuhn & Reiser, 2005; McNeill & Krajcik, 2007; Sandoval, 2003; Sandoval & Millwood, 2005). However, our results also build on this previous research in that we broke down the difficulty of this skill into a meaningful progression. Specifically, it is more difficult for students to *compare and critique* arguments based on the relevance and support of the evidence compared to simply *critiquing* within one argument or *identifying* new relevant-supporting evidence that is appropriate to include in an argument. Moreover, while it is more difficult for students to *critique* the evidence based on relevance and support within one argument than *identify* new

relevant-supporting evidence that is appropriate to include in an argument. While this progression posits critique as an important area for future k-12 argumentation instruction, it also suggests that identification may be a noteworthy intermediary step. It is our goal that teachers can use this progression to inform their instruction.

Informing Instruction

Formative assessments, such as the one presented in this study, are one way in which the scientific education research community can support teachers with the practice of argumentation (Osborne, 2010). Teachers can use the assessments to identify specific challenges of individual students or common student misundertandings (Furtak, 2012). In turn, teachers can use information gained from assessment results to inform future instruction based on their students' needs (Gotwals & Songer, 2010). For instance, if the better part of a class was able to correctly answer the identifies questions, but were not able to answer either critique question, then the teacher should target instruction around helping the students understand that high quality scientific evidence is relevant-supporting whereas low quality evidence is non-supporting or irrelevant. In current work, we are linking instructional strategies to each level of the construct map in order to better support teachers with moving their students along this progression.

The progression, however, can also be used preemptively to design instructional units based on the logical development of this concept (e.g., critiquing relevant-supporting evidence) (Furtak, 2012). This logical development could happen either within a single unit or across units in a year depending on the teacher's instructional goals. For instance, one teacher may want to incorporate students' learning how to critique relevant-supporting evidence in a unit he currently teaches on populations and ecosystems. He could incorporate in this unit one lesson each that

focus on identifying, critiquing, and comparing and critiquing relevant-supporting evidence. For instance, he could use a card sort (on a populations and ecosystems topic) in which students are asked to categorize relevant-supporting, relevant-contradictory, and irrelevant evidence cards when provided an investigation question and data that can be used to answer the question. Later in the unit, after the students have individually written an argument using data collected from an investigation they completed on populations or ecosystems, the teacher could incorporate a peer critique. Specifically, the students could switch papers with a partner and they could use a checklist to score the quality of their partners' evidence. Lastly, towards the end of the unit he could design a lesson in which students read arguments written by scientists on the topic of populations and ecosystems in which each presents a different perspective. They could then critique these arguments based on the relevance and support of the evidence, and ultimately choose with which perspective they most agree. In comparison, another teacher may decide to support students' learning how to critique scientific evidence across the school year. In this case, she might incorporate identifying relevant-supporting evidence into her unit on sound, critiquing relevant-supporting evidence on her unit on light, and comparing and critiquing relevant-supporting evidence on her unit structures of life. Regardless, whether preemptively or retrospectively, teachers as well as curriculum developers can use this meaningful progression to inform their instruction.

Supporting Students Across Modalities

The progression presented in this paper is for reading relevant-supporting evidence, which begs another question: *Should we be supporting students differently in different modalities?* The progression of students' abilities within this construct is mapped very

differently for the modality of writing. Otherwise stated, the construct is the same, however the construct maps differ by modality. Likewise, a students' ability within the same construct may not be equivalent in the two modalities. Specifically, several studies suggest that there are differences between the quality of students' oral and written arguments (Berland & McNeill, 2012; Knight & McNeill, in preparation; Sampson et al., 2010). As such, if there are differences (as we are suggesting), it is reasonable that the supports should vary depending on the modality.

Limitation and Future Work

We recognize that while this is an important first step in conceptualizing reading relevant-supporting evidence, the construct could be further expanded. First, the items of lowest difficulty are still above the ability level of some students. As previously discussed, our theorized lowest level (locate) was not the most appropriate lowest level. Yet, we believe there to be another level below identifying relevant-supporting evidence. In future work, we plan to pilot items that focus on identifying relevant evidence. Second, based on pilot results, we believe students' abilities to critique relevant-supporting, non-supporting, and irrelevant may not be equivalent. We would like to systematically tease these out.

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.

ATS (2007). *Reliability*. Retrieved from the American Thoracic Society webpage: http://qol.thoracic.org/sections/measurement-properties/reliability.html

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

Berland, L. K. & McNeill, K. L. (2012). For whom is argumen and explanation a necessary distinction? A response to Osborne and Petterson. *Science Education, 95*(5), 808-813.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement,* 12, 261-280.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cattell, R. B. (1966). The scree test for number of factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Cervetti, G., Perason, P. D, Bravo, M. A., & Barber, J. (2006). Reading and Writing in the Service of Inquiry-Based Science. In R. Douglas, M. Klentschy, & K. Worth (Eds.), *Linking Science and Literacy in the K-8 Classroom*. National Science Teachers Association.

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

Costello, A. B. & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation*, 10(7), 1-9.

Daultrey, S. (1976). *Principal components analysis*. Norwich: Geo Abstracts Limited.

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

du Toit, M. (2003). *Irt from SSI: Bilog-mg, multilog, parscale, testfact*. Lincolnwood, IL: Scientific Software International.

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

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.

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development 6*(2-3), 74-96.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Earlbaum Associates, Mahwah, NJ.
Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Education, 49*(9) 1181-1210.
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.
Green, K. E. (1996). Dimensionality analysis of complex data. *Structural Equation Modeling*, *3*, 50-61.
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.
Kim, J. O., & Mueller, C. W. (Eds.). (1978). *Factor analysis: Statistical methods and practical issues* (Vol. 14). Sage.
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 preparation). 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.
Lambert, Z. V., & Durand, R. M. (1975). Some precautions in using canonical analysis. *Journal of Marketing Research*, 468-475.
Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.
Lietz, P. (1995*). Changes in reading comprehension across cultures and over time*. Münster, New York: Waxmann.
Linacre, J. M. (2012). *A user's guide to Winsteps: Ministep Rasch-model computer programs* (v 3.74.0). winsteps.com
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
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.
Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, 87(2), 224-240.
National Research Council. (1996). National science education standards. Washington, D.C.: National Academy Press.
National Research Council (2012). *A framework for K-12 science education: Practices, Crosscutting Concepts, and core ideas*. Washington, DC: National Academy of Sciences.
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. 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.
Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for

matter and the atomic molecular theory. Measurement: *Interdisciplinary Research and Perspectives, 4*(12), 1-98.
Smith, R. M., Schumacker, R. E., & Bush, M.J. (1998). Using item mean squares to evaluate fit to the Rasch Model. *Journal of Outcome Measurement*, 2, 66-78.
Spivey, N. N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Resarch Quarterly*, 24(1), 7-26.
Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson/Allyn & Bacon.
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.
Wright, B. D., & Stone, M. H. (1979). *Best test design* (p. xiii). Chicago: Mesa Press.
Wu, M. L., Adams, R. J., and Wilson, M. R. (1997). ConQuest: Multi-Aspect Test Software, [computer program] Camberwell: Australian Council for Educational Research.

Appendix A: Reading relevant-supporting evidence items

Mrs. Warren asks her students to write an argument about the following question: **What is related to the height of ash clouds that are released from volcanoes?** Ben used the data table below to write his argument:

Name of Volcano	**Explosive Power of Volcano** (Scale: 0 to 8)	**Height of Ash Cloud**	**Wind Speed** at time of volcanic eruption
Kilauea	0	Very Low	12 mph
Galeras	2	Low	2 mph
Agung	4	Medium	18 mph
Santa Maria	6	High	5 mph
Tambora	7	Very High	7 mph

Ben's Argument:

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

(S1) Volcanoes that have more explosive power usually produce ash clouds that reach higher into the sky. (S2) The Tambora Volcano had an explosive power of 7, and the ash cloud was very high, and the Galeras Volcano had an explosive power of 2, and its ash cloud was low to the ground. (S3) The Tambora eruption was more powerful than the Galeras eruption, and this forced the volcanic ash from the Tambora Volcano much higher into the sky.

1. In which statement does Ben support his argument with evidence?
   a. Sentence 1 (S1) only
   b. Sentence 2 (S2) only
   c. Sentence 1 (S1) and Sentence 2 (S2)
   d. None

2. Ben is thinking of adding more evidence to his argument. Which piece of evidence best supports his claim?
   a. When the Eyja Volcano in Iceland erupted in 2010, the ash cloud was so large that it stopped planes from flying in parts of Europe for 6 days.
   b. The volcano called El Chichon erupted in 1982 with an explosive power of 4, and its ash cloud reached a medium height.
   c. The volcano called Abatar erupted in 1977 with an explosive power of 2, and its ash cloud reached very high into the sky.
   d. The most powerful volcanic eruption measured in the last 4,000 years was the 1815 Tambora eruption, which had an explosive power of 7.

3. Ben claims that volcanoes that have more explosive power usually produce ash clouds that reach higher into the sky. Ben wants to add the following piece of evidence:

   The volcano called Katami erupted in 1912 with a power of 6, and its ash cloud reached higher into the sky.

   This piece of evidence is…

   a. poor because it has nothing to do with Ben's claim.
   b. poor because it supports the opposite of Ben's claim.
   c. good because it supports a different claim than Ben's.
   d. good because it supports Ben's claim.

4. Anna is also in Mrs. Warren's class. Mrs. Warren asked Ben and Anna to compare arguments to see who used stronger evidence.

   Ben's Argument:

   Volcanoes that have more explosive power usually produce ash clouds that reach higher into the sky. The Tambora Volcano had an explosive power of 7, and its ash cloud was very high, and the Galeras Volcano had an explosive power of 2,and its ash cloud was low to the ground. The Tambora eruption was more powerful than the Galeras eruption, and this forced the volcanic ash from the Tambora Volcano much higher into the sky.

   Anna's Argument:

   Volcanoes that have more explosive power usually produce ash clouds that reach higher into the sky. The city of Pompeii was buried under about 15 feet of ash when the Mount Vesuvius Volcano erupted 1,934 years ago. Because the ash buried the city, it also preserved everyday items like animal bones, broken pieces of pottery, plants, buildings, and even art, which scientists use to learn about the lives of the people who once lived in Pompeii.

   Which student, Ben or Anna, better supports his or her argument? Why?

Appendix B: Reading Relevant-supporting Evidence Rubric:
Level 4: Compares and Critiques

Relevant-supporting evidence (RSE) General Rubric:

Score	Description
3 Correct Choice; Compare and Critique	Student **makes a correct choice and critiques the quality of the evidence by comparing the evidence** used in both arguments (relevant-supporting, relevant-contradictory, or irrelevant).
2 Correct Choice; Critique	Student makes a correct choice and critiques the quality of the evidence used in one of the arguments (relevant-supporting, relevant-contradictory, or irrelevant).
1 Correct Choice	Student makes a correct choice and writes that one argument provides stronger evidence. However, the identified reason **is wrong.** OR Student **does not identify any reason** even though s/he makes a correct choice
0 No Choice; Incorrect Choice	Student does **not make any choice** OR Student **makes a wrong choice**

Relevant-supporting Evidence (RSE) Specific Rubric

Score	Description	Student Response
3 Correct Choice; Compare and Critique	Student **makes a correct choice and critiques the quality of the evidence by comparing the evidence** used in both arguments (relevant-supporting, relevant-contradictory, or irrelevant).	RE1AT8S10 *Ben's argument was by far much better than Anna's. Ben's argument stated the explosive power and height at the ash cloud while supporting the claim. Anna's argument did none of that. Also, Anna's argument seemed to travel to a whole different topic. She did explain the power of Mt. Vesuvius but there was no definitive evidence that the higher the power, the higher the ash cloud. It got to a point where she was talking about history not science. Therefore, Ben's argument was the best.* [Student identifies the RS variables in Ben's argument and identifies Anna's evidence as irrelevant]. RE1BT18S12 *Terrance because Dina's argument talked about a time when there were more than usual earthquakes that release that many tons when he is trying to say that there are less earthquakes that release that much energy. In Terrance's argument he compares how many earthquakes in the year that are bigger and smaller to show there are way more smaller earthquakes than there are big.* [Student says Tina's evidence is RC, and Terrance's is RS]. RE1CT18S8 *Shawn better supports his argument. He uses evidence to support the claim. Alison however, used evidence that did not support the claim. She said cool magma can be thin and runny, but it's the complete opposite of the claim. Shawn sticks to the claim and uses excellent statistics.* [Student identifies that Shawn's argument had RSE and Alison's was RCE]. RE2BT9 *Sarah better supports here augment because she uses evidence to support her argument and she explains how earthquakes happen and why longer earthquakes are more destructive. Teddy's argument is poor because has given evidence that doesn't support his claim, but goes against it. He has also included an explanation of earthquakes, but he uses it in an incorrect way.* [Student identifies that Sarah's had RSE and Teddy's was RCE] **It is acceptable to point to the data in the argument as opposed to using the terminology. For instance, if the student identifies the relevant-contradictory evidence and says that it is bad, then that is acceptable.** *Terrance's argument was better because it talked about multiple years instead of just 2011, like Dina. Plus, how Dina supported his argument wasn't very good because he said that usually there are only 1319 earthquakes that release that much energy, instead of 2276.* **If they say. "this argument has evidence" it implies that it is relevant-supporting; however if they use words such as "data", "supporting details", "answered the question", "example" etc., instead of "evidence," they have to explain/clarify what they mean in order to be given credit for RSE.** *RE1AT18S7 I feel Ben answered the question better. He answered the question and gave supporting details. June though stated the question but didn't answer it well enough and I felt she got a little of topic by going into to much detail about the remains of Pompeii. Ben showed differences between explosive power and how high the two volcanoes went while June except for the first sentence didn't mention explosive power.* **"About the subject" or "stays on topic" is not enough because it is not necessarily supporting (although relevant).**

		"Better proof" without reason is not sufficient to imply relevant-supporting.
2 Correct Choice; Critique	Student makes a correct choice and critiques the quality of the evidence used in one of the arguments (relevant-supporting, relevant-contradictory, or irrelevant).	RE1AT18S13 *Ben supports his argument better. He does because Anna went off-topic by talking about Pompeii's.* [Doesn't say why Ben's is better; only that Anna's was irrelevant]. RE1AT18S2 *Ben's argument supports his argument because he states what he needs. Ben states the power of the volcano, then says how high the ash went. It shows what Ben claims better than Anna's.* [Only states the RS variables in Ben's argument; doesn't explain what was bad about Anna's evidence]. RE1AT18S2 *Terrance supports her argument better because she state's different earthquakes. She say that earthquakes with less energy happen more often than earthquakes with more energy. She shows all the evidence she needs.* [States the RS variables in Terrance's argument; Doesn't mention why Dina's evidence is bad]. RE1AT18S15 *Shawn's argument is better. He has everything you need to make a great argument. He has good evidence, data, and it all ties in to the question. Alison does that, but not as thoroughly as Shawn.* [Student identifies that Shawn's argument as good evidence and because it is evidence it is RS, but they are wrong about Alison's argument]. RE2BT9S8 *Sarah's argument is better than Teddy's because he addresses the question but fails to support the claim. For instance, in sentence 2 of Teddy's paragraph he failed to support the claim.* [Student identifies the Sarah's is better and says it is because Teddy's evidence is non-supporting; does not critique Sarah's evidence.] **It is acceptable to point to the data in the argument as opposed to using the terminology. For instance, if the student identifies the relevant-contradictory evidence and says that it is bad, then that is acceptable.** *Terrance's argument was better because it talked about multiple years instead of just 2011, like Winston. Plus, how Dina supported his argument wasn't very good because he said that usually there are only 1319 earthquakes that release that much energy, instead of 2276.* **If they say. "this argument has evidence" it implies that it is relevant-supporting; however if they use words such as "data", "supporting details", "answered the question", "example" etc., instead of "evidence," they have to explain/clarify what they mean in order to be given credit for RSE.** **"About the subject" or "stays on topic" is not enough because it is not necessarily supporting (although relevant).** **"Better proof" without reason is not sufficient to imply relevant-supporting.**

1 Correct Choice	Student makes a correct choice and writes that one argument provides stronger evidence. However, the identified reason **is wrong.** Student **does not identify any reason** even though s/he makes a correct choice	RE1AT18S5 *I think Ben's argument was better. Ben explained more facts about the topic. Anna just gave an example; she needed facts like Ben had. Ben had more facts about volcanoes. That is why I think Ben had the better argument.* [Does not clearly define Ben's as RSE or Anna's as irrelevant]. RE1BTS15 *I think Terrance's argument better supports his argument because he explain what he was stating thoroughly. He told us that if "less earthquakes happen when lots of energy is released". He backed up his answer with detains that support what he was thinking. I also feel that his writing was organized and was on topic the whole time so the teacher or reader/audience could understand what point he was making. Terrance said that there are many earthquakes that don't release a lot of energy like the ones the don't happen often.* [Correct choice, but the reason is not clear enough to count]. RE1AT18S5 *I think Shawn's argument was better because there were facts stated. She gave 1 example but didn't go on and on about it! Shawn got right to the point with good facts.* [Correct choice, but the reason is not clear enough to count]. RE1BT18S4 *Shawn's argument was better b/c Angie put information that isn't current.* [Correct choice, but the reason is not clear enough to count]. RE1CT18S10 *Shawn had a better argument. He supports his argument by stating examples. He also states the question in the answer.* [Correct choice, but the reason is not clear enough to count]. RE2AT9S3 *Sarah's argument because she gives two examples of earthquakes to compare* [More does not mean better; the critique is wrong.]
0 No Choice; Incorrect Choice	Student does **not make any choice** OR Student **makes a wrong choice**	RE1BT18S3 *I think Junes argument is better because she uses a known example like Pompeii which gives you a better understanding about what she is talking about. [Wrong Choice].* RE1AT18S5 *I think that Dina's argument was better. I think Dina's was better than Terrance's argument because Dina has the number of earthquakes in one year and it was well explained. On the other hand Terrance had facts but he didn't explain it as much as Winston did. [Wrong Choice].* *I think Alison's argument was better because it has details. It had information that could help me understand what I was reading. [Wrong Choice].* RE1AT18S3 *I think Teddy's is better because she has more evidence to support the argument. [Wrong Choice].*