Measuring Students' Scientific Argumentation Associated with Uncertain Current Science

Hee-Sun Lee[1]

Amy Pallant[2]

Sarah Pryputniewicz[2]

Ou Lydia Liu[3]

[1] University of California, Santa Cruz

[2] The Concord Consortium, Presenter

[3] Educational Testing Service, Princeton

Strand 10: Curriculum, Evaluation, and Assessment
Related Paper Set: Assessing Scientific Argumentation: Challenges and Future Directions
8:30am-10:00am, San Cristobal, April 8, Monday, 2013

Lee, H. –S., Pallant, A., Pryputniewicz, S., & Liu, O. L. (2013). Measuring students' scientific argumentation associated with uncertain current science. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Rio Grande, Puerto Rico.

Abstract

In this study, we investigate how students' claim, justification, uncertainty qualifier, and conditions of rebuttal contribute to the measurement of their overall scientific argumentation ability. We designed six sets of items related to climate change and the search for life on other planets. The items were administered to 956 students taught by 12 middle and high school teachers in the Northeastern United States. Rasch analysis results based on the Partial Credit Model indicate that (1) students' responses to all four argumentation elements can be interpreted on a single scale, (2) higher scientific argumentation abilities are needed in the order of uncertainty qualifier, claim, justification, and conditions of rebuttal on the scientific argumentation scale, (3) justification and conditions of rebuttal items measure a wider range of the scientific argumentation scale than claim and uncertainty items, (4) students who make a single warrant are more likely to think about conditions of rebuttal within the context of investigation, and (5) students who make two or more warrants are more likely to consider conditions of rebuttal beyond the context of investigation. We discuss these results to refine Toulmin's theory, provide recommendations for designing and implementing large scale assessment, and suggest future research directions.

Key words: Scientific argumentation, construct modeling, Rasch Analysis

Introduction

To make science learning authentic to scientists' practices and meaningful to students' everyday lives, scientific inquiry has been advocated (National Research Council, 1996; Schwab, 1962). The process of scientific inquiry starts with a driving question, ensues with an investigation, and concludes with a claim based on evidence collected from the investigation (Koslowski, 1996; Latour & Woolgar, 1985). Since the culminating step in scientific inquiry is communicating with others about findings from an investigation (Bricker & Bell, 2008), scientific argumentation has been considered a critical element of inquiry-based science curriculum, instruction, assessment, professional development, and learning environment (Berland & McNeill, 2010; Duschl & Osborne, 2002; Duschl, Schweingruber, & Shouse, 2007; Jimenez-Aleixandre, Rodriguez, Duschl, 1999; Kuhn, 1993; Lawson, 2003; McDonald, 2010; McNeill & Pimentel, 2010; Yerrick, 2000; Zembal-Saul, 2009; Zohar & Nemet, 2002). As a result, research on scientific argumentation has surged in the last decade (Lee, Wu, & Chai, 2008), leading to numerous frameworks to analyze rhetorical and dialogic arguments generated in the science classroom (Clark, Sampson, Weinberger, & Erkens, 2007; Sampson & Clark, 2008).

Scientific argumentation consists of claim and justification and can happen in either rhetorical or dialogic form. Toulmin (1958) specified that a rhetorical argument may include up to six elements such as claim, data, warrant, backing, modal qualifier, and conditions of rebuttal. Guided by Toulmin's classification, science education research has been undertaken to analyze scientific arguments in students' essays (Kelly & Takago, 2002), lab reports (Erduran, Simon, & Osborne, 2004), tests (Zohar & Nemet, 2002), classroom discourse (Chinn & Osborne, 2010; Hogan, Nastasi, & Pressley, 2000; Kuhn, 2010; McNeill & Pimentel, 2010), and online discourse (Sampson & Clark, 2009). In these studies, claim, data as evidence, and warrant and backing as

justification were the most consistently analyzed argument elements. Modal qualifiers and conditions of rebuttal were less systematically studied in part because they did not frequently occur in students' naturalistic discourse or written artifacts without specific prompts (Schwartz, Neuman, Julia, & Llya, 2003). The few studies that investigated these two elements focused on dialogic discourse involving multiple parties defending different claims (Kuhn, Shaw, & Felton, 1997). In that context, conditions of rebuttal were considered as qualifiers (Clark et al., 2007) and rebuttals as counterarguments (Erduran et al., 2004; Means & Voss, 1996). Modal qualifiers that Toulmin (1958) defined as "the strength conferred by the warrant" (p. 101), e.g., "necessarily," "frequently," and "most likely," were largely overlooked even though few scientific claims and justifications are made with absolute certainty due to incomplete or insensitive measurements, limitations in current theory or model, and complexity involved in phenomena under investigation (American Association for the Advancement of Science, 1993).

Though a lot of attention has been paid to identifying student performance levels on student-generated arguments, the use of the currently available analytic frameworks at the large scale is limited. For the large scale assessment purpose, a more parsimonious construct is needed because it is not feasible to compare students separately on claim, justification, qualifier, and conditions of rebuttal in state or national testing. Most analytical frameworks enabled researchers to tally frequencies in each argument element. Occasionally, scores on claim, data, and reasoning were added together, even if scores in these argumentation elements may not be on the same interval scales. Therefore, the purpose of this study is to investigate an analytic framework for establishing students' overall argumentation ability on a single scale.

In this study, we characterize scientific argumentation as a multi-level construct based on students' claims, justifications for their claims, uncertainty qualifiers, and conditions of rebuttal.

We designed six item sets to elicit these argument elements related to climate change and life in space. The research questions of this study are:

- What types of claims, justifications, uncertainty qualifiers, and conditions of rebuttal do students provide when they formulate rhetorical scientific arguments?
- How are students' claims, justifications, uncertainty qualifiers, and conditions of rebuttal mapped onto the underlying scientific argumentation construct?

We first summarize literature related to scientific argumentation and sources of uncertainty. Next, we introduce a scientific argumentation construct map and describe research methods related to instrument design, subjects, and data collection and analysis procedures. We present and discuss results in the order of research questions listed above, followed by implications for science teaching and science education research.

Literature Review

Argument

Though argument and argumentation are interchangeably used in the literature without clear distinction, we use argument throughout this paper to mean reasoning or justification to support an assertion or conclusion (Zohar & Nemet, 2002) and argumentation as a skill or ability associated with formulating arguments. Kuhn and Udell (2003) differentiated dialogic or dialectical arguments from rhetorical arguments constructed by individuals as saying "two or more people engage in debate of opposing claims" (p. 1245). Another form is analytical argument based on pure logic (van Eemeren et al., 1996). Argument is recognized as a process and as a product (Berland & McNeill, 2010). Argument is a verbal, social, and rational activity. As arguments occur across disciplines, Toulmin (1958) extracted six field-invariant argument elements (See Figure 1):

- Claim (C) or conclusion "whose merits we are seeking to establish" (p. 97)

- Data (D) are "the facts we appeal to as a foundation for the claim" (p. 97)
- Warrants (W) "show that, taking these data as a starting point, the step to the original claim or conclusion is an appropriate and legitimate one" (p. 98)
- Modal qualifiers (Q) indicate "the strength conferred by the warrant" (p. 101) and "some warrants authorize us to accept a claim unequivocally with the adverb 'necessarily' and others authorize us to make the step from data to conclusion either tentatively, or else subject to conditions, exceptions, or qualifications-in these cases other model qualifiers such as 'probably' and 'presumably' are in place" (pp.100-101)
- Conditions of rebuttal (R) indicate "circumstances in which the general authority of the warrant would have to be set aside…exceptional conditions which might be capable of defeating or rebutting the warranted conclusion" (p.101) and are directly connected to the choice of the modal qualifier.
- Backing (B) shows "assurances without which the warrants themselves would possess neither authority nor currency" (p.103).

---------- Insert Figure 1 Here ----------

In science education research, rebuttals have largely been attributed to counterarguments during classroom discourse (Kuhn, 2010), group argument construction (Osborne, Erduran, & Simon, 2004), and online discussion (Samson & Clark, 2009). A few studies characterized qualifiers as "special conditions under which the claim holds true" (Clark & Sampson, 2007, p.347), rather than the original Toulmin's description of "some explicit reference to the degree of force which our data confer on our claim in virtue of our warrant" (p.101) such as "presumably," "always," and "almost certainly." Therefore, current uses of qualifiers in the scientific education community resemble conditions of rebuttal in Toulmin's terminology.

Toulmin (1958) pointed out that, though these argument elements are field invariant, backing provides field-dependency for "the criteria or sorts of ground required to justify" a claim (p. 36). The scientific knowledge base along with the established and accepted scientific inquiry methods provides sources of backing needed in formulating and evaluating scientific arguments.

Uncertainty and Conditions of Rebuttal in Scientific Argument

A frequently utilized modal qualifier in scientific arguments by the community of scientists is uncertainty. Uncertainty is associated with one's confidence or lack thereof in describing current phenomena or predicting outcomes. Uncertainty occurs because the knowledge, experience or information used in descriptions or predictions is not sufficient enough to provide definite and exact claims.

Scientific uncertainty. Any scientific claim involves uncertainty to some extent. Scientific uncertainty is related to conceptual and methodological limitations imposed by the particular scientific inquiry method applied to an investigation. Scientific uncertainty associated with measurement, probability, phenomena, and status of current knowledge base can weaken the strength of an argument thus can be subject to rebuttal.

- Measurement uncertainty: Measurement is a "process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity" (Joint Committee for Guides in Metrology, 2008, p. 16). Even though a quantity such as the distance between the Sun and the Earth is considered to have a true quantity value, an instrument designed to measure it may not produce the true quantity value. The difference between the measured and the true quantity values is called measurement error. To reduce the measurement error, the same quantity is measured multiple times. The standardized error of measurement indicates the degree of uncertainty associated with the measurement of

the quantity. In addition, measurement uncertainty can arise rather systematically due to the accuracy, precision, and resolution of a measuring instrument.

- Probabilistic uncertainty: Scientific claims expressed in probability show mathematical uncertainty. Probability describes the likelihood of a certain event to occur, such as having a 60% chance of a rain shower. Using probability distributions, all possible events are considered while none of the events are completely ruled out. Probability has been used in a variety of disciplines to address uncertainty especially in describing molecular, atomic, and subatomic phenomena as well as in predicting natural disasters such as hurricanes and earthquakes.
- Uncertain phenomena: Some scientific phenomena under investigation can be uncertain. The best example is Heisenberg's uncertainty principle, which states that the position and the momentum of a particle cannot be known to an observer with the same accuracy at the same time (Lindley, 2007). This is considered the characteristic of the scientific phenomenon itself, not the fault of the measurement method. Moreover, most scientific phenomena are complex because they involve an extremely large number of entities whose interactions are governed by numerous known and unknown factors over extremely short or long periods of time. In these cases, scientific claims cannot be stated with absolute certitude because of unexamined elements of a phenomenon in a study.
- Uncertainty due to collective understanding at the time: The knowledge, equipment, tools, and driving questions currently employed by scientists limit their claims and explanations. This uncertain nature of science enterprise is captured as "tentativeness" of scientific knowledge in studies of nature of science (Lederman & O'Malley, 1990). For example, on the 125th anniversary of its publication, *Science* magazine selected 125 topics that current science cannot answer but "scientists should have a good shot at

answering the questions over the next 25 years, or they should at least know how to go about answering them" (Kennedy & Norman, 2005, p. 75). Among the questions was "Are we alone in the Universe?" Since our understanding of life is very much limited to life on Earth and at the same time the Universe is vast, our theoretical and empirical tools of finding extraterrestrial life are very much limited.

Students' uncertainty. Unlike scientific uncertainty, in which the limitations reflect the current status of scientific knowledge and investigation methods commonly adopted by the community of scientists, student uncertainty can reflect the student's assessment of his own status of knowledge, ability, and skill. This aspect of uncertainty is related to self-efficacy, referred to as individual students' judgments of their capabilities to perform a given task (Bandura, 1986). Among the self-concepts that students hold about themselves, self-efficacy is most profoundly related to students' academic performances (Pintrich & De Groot, 1990). Students are continuously learning science through adding, comparing, contrasting, and revising various ideas from their own experience and from informal and formal science education (Linn & Eylon, 2006). When students are asked to explain their claim in response to a scientific question, it is likely that they do not hold a tight grasp of knowledge and experience required to answer the question. In dealing with ambiguous or anomalous evidence, students often attempt to alter the evidence to fit the scientific disposition they are set out to prove (Chinn & Brewer, 1993; Germann & Aram, 1996). Metz (2004) discovered five spheres of elementary school students' uncertainty related to how to produce a desired outcome, data, trend identified in the data, generalizability of the trend, and the theory that can explain the trend, indicating that students can consider uncertainty as part of inquiry-based investigations.

Analyzing Students' Scientific Arguments

Most frameworks designed to analyze rhetorical or dialogic arguments distinguished well- from poorly-constructed arguments. Increasing competence has been identified in justifications, conditions of rebuttal, and counterarguments as follows.

Justifications. Though students make arguments in everyday life and appear to be doing so naturally (Simosi, 2003), they are not inclined to make arguments in science class. Often, students do not include justifications for their claims (Bell & Linn, 2000; Sandoval & Millwood, 2005). Justifications show how students coordinate data or evidence with claims (Duschl & Osborne, 2002). Selecting salient evidence from all available data is considered important (McNeill, Lizott, Krajcik, & Marx, 2006). Students' ability to justify is strongly correlated with students' knowledge of science content (Means & Voss, 1996). However, Ohlsson (1992) pointed out that having knowledge cannot guarantee its use because "theory does not prescribe its own articulation" (p. 183). Instead, the student needs to actively apply a theory "to a particular situation, to decide how exactly, the theory should be mapped onto that situation, and to derive what the theory implies or says about that situation" (Ohlsson, 1992, p. 182). Coding rubrics for justifications focused on whether and how many scientifically-valid justifications were included. For instance, Clark and Sampson (2008) coded for the grounds students used in the order of (1) claim only without grounds, (2) data only, and (3) multiple data or justified data. Means and Voss (1996) counted the number of reasons. Zohar and Nemet (2002) counted the number of justifications in three scoring categories: no scientifically-valid justifications (score 0), one valid justification (score 1), and two or more valid justifications (score 2).

Conditions of rebuttal. Walton, Reed, and Macagno (2008) proposed three types of rebutting an argument. The first type is to argue that premises, i.e. data in Toulmin's terminology, used in an argument are not true. The second type is to argue that "the conclusion does not follow from the premises" (p. 222), indicating faults in warrants or backing. The third

type is to argue that "the conclusion is false, or at any rate, that there are reasons to think so" (p. 222), leading to a counterargument. Walton et al. (2008) distinguished between rebuttal and refutation where the former is to simply oppose an argument while the latter not only is opposed to the original argument, but also has enough strength to overpower it. Conditions of rebuttal written in an argument are mainly of the first type when they are presented without counterclaims or counterarguments. Students who realize claims as being conditional and elaborate how and in what conditions claims can be limited are considered to have higher reasoning abilities than those who do not (Means & Voss, 1996).

Counterarguments. Counterarguments are an example of the third type of Walton et al. (2008)'s rebuttals, and they oppose other arguments with their own evidence, justifications, and claim. Analyses of counterarguments often occur in dialogic argument situations such as group or classroom discussions where opposite points of view are elicited and debated. Arguments that address potential counterarguments are considered more effective than without them (Erduran et al., 2004; Kuhn, 2010; McNeill & Pimentel, 2010; Sadler & Fowler, 2006).

Overall scientific argumentation ability. Most frameworks analyzed students' scientific arguments in multiple coding categories and compared frequencies of occurrences in each coding category. To represent overall performance on scientific argumentation, researchers used three methods. First, a new set of categories were created by adding scores from two or more argument categories. Erduran et al. (2004) used claims (C), data (D), warrants (W), backing (B), and rebuttals (R), to create CD, CW, CDW, CDR, CDWB, and CDWR. In these combinations, CDWB and CDWR represent a higher scientific argumentation performance than the other four. Since this method created categorical variables, only frequency comparisons were used in the analysis.

Second, a multi-level ordinal scale was created. Erduran et al. (2004) defined the first level as only claims or counter claims, the second level as claims with data, warrants, or backings. The third level added weak rebuttals to the second level while the fourth level added one clearly identifiable rebuttal. The fifth level included multiple clearly-identifiable rebuttals. Osborne et al. (2004) used this ordinal scale to characterize to what level a dialogic argument situation could reach. Sadler and Fowler (2006) developed a five-point argumentation quality rubric consisting of claims without justifications, with no valid grounds, simple grounds, elaborated grounds, and elaborated grounds with a counter-position. Sadler and Fowler (2006) applied multivariate analyses of variance on this argumentation quality variable.

Third, a total score was created by summing up scores each student received on multiple coding categories. For example, after giving a point for each of claim, data/evidence, reasons and backing, qualifier to construct an argument, counterargument, and rebuttals, Chinn and Osborne (2010) used a composite score to find relationships between students' scientific argumentation ability and instructional practices. Zohar and Nemet (2002) combined the number of justifications scored 0 to 2 with the argument structure scored 0 (no valid justification), 1 (a claim supported by a justification) and 2 (a claim supported by multiple justifications with multiple conditions of rebuttal). Similarly, Sampson and Clark (2009) added scores assigned to explanation sufficiency, conceptual quality, evidence quality, and reasoning adequacy categories to produce an overall argument score.

Summary

Many analytic frameworks have been developed and applied to students' rhetorical and dialogic arguments in the past decade. Despite variations among these frameworks, similar patterns emerged in distinguishing more from less competent responses within justifications, conditions of rebuttals, and counterarguments. However, student performance levels across these

argumentation variables were not yet accumulated to represent students' overall scientific argumentation ability on a psychometrically defensible scale.

Methods

We first define the scientific argumentation ability with a construct map (Wilson, 2004) which guided the instrument design, scoring, analysis, and interpretation in this study. We then describe research methods in a four-step process suggested by Mislevy and Ricoscente (2005): assessment task selection (instrument design), assessment task presentation to students (data collection), evidence identification that shows the assignment of designated scores within argumentation coding categories (scoring rubrics), and evidence accumulation that shows how student responses were amassed across multiple coding categories (scale development).

Rhetorical Scientific Argumentation Construct Map

Based on Toulmin's argument structure (1958), we conceptualize the scientific argumentation construct consisting of six distinct levels. Table 1 shows these levels on a continuum in the order of increasing sophistication. Higher levels are assigned to students who include more elements in their scientific arguments. The first level represents non-scientific statements. In the second level, students write or choose only a scientific claim without supporting evidence or relevant knowledge. In the third level, students make a claim based on salient data or relevant knowledge. In the fourth level, students make a claim based on coordination between salient evidence and relevant knowledge. In the fifth level, students modify the strength of their scientific argument. In the sixth and highest level, students can distinguish conditions where their scientific arguments are held true and recognize limitations associated with measurement, current knowledge base or model, and phenomena.

---------- Insert Table 1 Here ----------

Instrument Design

We selected two science topics, climate change and the search for life on other planets, from the 125 science problems a panel of scientists identified as "What We Don't Know" in *Science* (Kennedy & Norman, 2005). It is essential to use these current science topics in order to encourage students to elicit their uncertainty and conditions of rebuttal in their arguments. In cases where scientifically correct answers are obvious and superior, students' uncertainty might not be fully elicited. Three scientific investigations on the topic of climate change were used as item contexts:

- Pinatubo item set: describing how Mountain Pinatubo eruptions impacted global temperatures;
- T2050 item set: predicting the temperature of 2050 based on the ice core records of global temperatures and atmospheric CO_2 levels between 125,000 years ago and 2000;
- Ocean item set: predicting the trend of atmospheric CO_2 level when ocean temperature increases.

For the topic of life on other planets, three investigations were chosen:

- Galaxy item set: predicting a possibility of finding extraterrestrial life based on the number of galaxies and stars observed in the Universe;
- Life item set: predicting existence of Earth-like life forms by comparing information between an imaginary planet called Athena and the Earth;
- Spectra item set: predicting conditions between Uranus and Neptune based on absorption spectra.

For each of these six investigations, we stringed together four items consisting of making scientific claims (claim), explaining scientific claims based on evidence (justification), expressing the level of uncertainty about explanations for the claims (uncertainty), and

describing their source of uncertainty (conditions of rebuttal). We asked these elements separately since the use of qualifiers and the consideration of rebuttals do not naturally occur in students (Sandoval, 2003). For claims, either multiple-choice or short-answer item format was used. For justifications, we provided data in graphs, tables, or written statements and asked to "Explain your answer" in an open-ended format. Then, students were asked to rate their uncertainty on a five point Likert scale from "1" being not certain at all to "5" being very certain. Students were asked to explain their uncertainty. See Figure 2 for the Life item set. On the scientific argumentation construct, the claim items were designed to match the first two levels; the justification items from the first to fourth level; uncertainty items to the fifth level; the conditions of rebuttal items from the fifth to the sixth level. Since the items were answered by individual students, how students formulate counterarguments was not addressed in this study.

---------- Insert Figure 2 Here ----------

According to the scientific argumentation construct shown in Table 1, we hypothesized that higher and higher scientific argumentation abilities would be needed in the order of claim, explanation, uncertainty, conditions of rebuttal items for students to be successful.

Data Collection

We developed a test consisting of the six item sets. In the first month of a new school year, the test was administered online to a total of 956 Earth Science students taught by 12 teachers in six middle and high school schools located in the Northeastern United States. Among the students, 52% were female; 90% spoke English as their first language; 83% were middle school students; and 70% used computers regularly for homework. It took about 30 to 40 minutes for students to complete the test. We eliminated students who did not complete more than 50% of the 24 items to ensure the accuracy of the ability estimates. As a result, 837 students were included in the analysis. Average missing data percentages were 2.9% on claim and

justification items and 4.2% on uncertainty and conditions of rebuttal items. The missing data were treated as missing as Rasch analyses are tolerant of missing data.

Scoring Rubrics

Multiple-choice and short-answer claim items were dichotomously coded, "1" for claims that were consistent with what current scientists would claim and "0" for claims that were not. Explanation items were coded based on whether scientifically relevant evidence or relevant pieces of knowledge was included and how well students coordinated between knowledge and evidence. Figure 3 shows a scoring rubric for the justification item in the Life item set. Justifications without science-related information were assigned to the no evidence category (score 1). Students can use as many evidence pieces as possible from the data provided in the item or as many knowledge pieces as possible. When justifications included relevant data but did not include how or why the data supported their claims, they were assigned to the relevant evidence/knowledge category (score 2). When justifications included relevant knowledge without data, they were also assigned to the relevant evidence/knowledge category. Justifications that coordinated between a piece of salient evidence and a piece of relevant knowledge were assigned to the single warrant category (score 3). There were five possible coordinated links in the Life item set as shown in Figure 3. When justifications provided two or more coordinated links between evidence and knowledge, the two or more warrants category was assigned (score 4).

---------- Insert Figure 3 Here ----------

On uncertainty items, "1" and "2" responses were assigned to uncertain (score 0), "3" responses to neutral (score 1), and "4" and "5" responses to certain (score 2) categories. Student responses to conditions of rebuttal items were assigned to four levels as shown in Table 2. The first level (score 0) included blank, off-task responses, and restatements of claims or uncertainty

ratings. The second level (score 1) represented students' status of knowledge and ability related to the science topic addressed in the item. The third level (score 2) dealt with scientific uncertainty involved in the outcome, knowledge, and data related to the investigation addressed in the item set. The fourth level (score 3) represented scientific uncertainty beyond the investigation featured in the item.

---------- Insert Table 2 Here ----------

Data Analysis for Scale Development

We used descriptive statistics to show what types of scientific claims, justifications, uncertainty levels, and conditions of rebuttal students exhibited. Since we had claim items scored from 0 to 1, justification items from 0 to 4, uncertainty items from 0 to 2, and conditions of rebuttal items from 0 to 3, we used the Rasch Partial Credit Model (Rasch, 1966) shown below to fit the data (PCM; Wright & Masters, 1982):

$$P_{nix}(\theta) = \frac{\exp[\sum_{j=0}^{x}(\theta_n - \delta_i - \tau_{ij})]}{\sum_{r=0}^{m_i}[\exp\sum_{j=0}^{r}(\theta_n - \delta_i - \tau_{ij})]}$$

where $P_{nix}(\theta)$ stands for the probability of student *n* scoring *x* on item *i*. θ stands for the student location on the scientific argumentation construct in this study. δ_i refers to the item difficulty. τ_{ij} (j = 0, 1, ..m) is the step parameter indicating the difficulty of achieving each score *j* for item *i*.

We used the *ConQuest* software (Wu, Adams, & Wilson, 1997) to conduct a Rasch analysis based on the Partial Credit Model. ConQuest provides an ability estimate for each student and an item difficulty estimate for each item. Both ability and item estimates are calibrated to be on the same logit scale with values ranging from -4.0 to 4.0. The higher the logit

value, the more able the student and the more difficult the item. We used fit statistics to examine whether observed student responses to claim, justification, uncertainty, and conditions of rebuttal items fit the partial credit model. We then examined overall item difficulties to determine how these four argumentation elements can be ranked according to the level of ability required on the scientific argumentation scale. We also used item-test consistency to show how well each item performed as compared to the entire test. We used the Wright Map to compare the distributions of student abilities and item thresholds on the scientific argumentation scale. An item threshold indicates how difficult it is for students to achieve a designated score within each item.

Results

Student Response Distributions

Table 3 shows how students' responses were distributed across the six item sets in terms of claim, justification, uncertainty, and conditions of rebuttal.

Claims. Overall, 49.6% of the students' claims were scientific. Scientific claims related to the T2050, Ocean, and Spectra item sets occurred much less frequently than in the other three item sets. The lower scientific claim rates for the T2050 and Ocean items were related to students' difficulty with interpreting graphical representations that did not provide direct answers and writing open-ended claims (note that the other four claims were multiple-choice claim items). For example, the T2050 item context featured two graphs: global temperatures and atmospheric CO_2 concentration levels over a 125,000 year period. The Ocean item featured the solubility of CO_2 in ocean water over a range of temperatures while students predicted what would happen to the atmospheric CO_2 level if the ocean temperature increases. The Spectra item's claim was difficult because most students did not learn absorption spectral lines of the light reflected on Neptune and Uranus prior to the testing.

---------- Insert Table 3 Here ----------

Justifications. About half of the justifications did not include any scientifically relevant evidence or knowledge while slightly more than one third included either salient evidence or relevant knowledge for the claim. The coordination between evidence and knowledge as shown in warrants was difficult to achieve as only 13.1% of the responses included a scientifically elaborated warrant and 2.3% included two or more elaborated warrants. Students' justification levels were relatively lower in the T2050 and Ocean items in which students also had difficulty in making scientific claims. On the Pinatubo item, 59.5% of the students were able to pinpoint the evidence related to the global temperature decline resulting from volcanic eruption (Relevant evidence). However, most students did not explain how volcanic eruption would cause the global temperature to drop. Interestingly, students tended to formulate warrants more effectively with the Galaxy and Life items. Current science cannot provide definite claims related to whether life exists outside of Earth based on a limited set of data. This indicates that students could be more willingly engaged with scientific argumentation when science is uncertain.

Uncertainty. More than half of students' overall uncertainty ratings indicate that they were certain about their arguments. Two thirds of students were certain about their arguments in the Pinatubo, Ocean, Galaxy, and Life items. In contrast, students were very uncertain about their arguments in the T2050 and Spectral items. Even though most students could not write a scientifically correct claim or elaborated warrants, they were certain about their argument in the Ocean item set, indicating that students attempted to find a direct answer from the graph shown. Apparently, students did not differentiate the solubility of CO_2 from the level of atmospheric CO_2, leading to the opposite claim.

Conditions of rebuttal. Overall, 40.1% of students' responses did not indicate what made their arguments certain or uncertain. The most predominant conditions of rebuttal were whether students were able to understand the question, the related science knowledge, or the data

provided in the item. In some cases, students relied on authorities such as books, news, and teachers. Only 15.0% of the student responses mentioned scientific uncertainty related to the data and the knowledge relevant in the investigation. Very few responses (2.7%) went beyond the investigations featured in the items. In the example of the Life item set, students provided some issues that might undermine their arguments such as "Maybe a different form of life that is not affected by UV rays, extreme heat, and low oxygen is on that planet, like bacteria *[limitation on the current knowledge of life]*."

Rasch Scale for the Scientific Argumentation Construct.

Item fit. Table 4 shows item fit statistics in mean square values. The acceptable range for item fit to the Rasch Partial Credit Model is between 0.70 and 1.30 (Bond & Fox, 2007). There were no misfit items based on infit and outfit statistics. According to these results, students' responses to all four types of items could be interpreted on the overall scientific argumentation scale.

---------- Insert Table 4 Here ----------

Item difficulty. We examined how difficult each item was on the scientific argumentation scale. Table 4 shows that the easiest item on the scale was the claim item in the Life item set with the item difficulty value of -2.24. This means that students whose scientific argumentaiton ability were at -2.24 had a 50% chance of answering this item correctly. The most difficult item was the claim item in the Ocean item set with the item difficulty value of 1.16. We then compared average item difficulty values across claim, justification, uncertainty, and conditions of rebuttal items. The easiest item group was uncertainty items, followed by claim item group. The most difficult item group was conditions of rebuttal. Justification items were placed between claims and conditions of rebuttal. See Table 4. These results indicate that the order of the required ability on the scientific argumentation scale was uncertainty → claim → justification →

conditions of rebuttal, instead of the hypothesized order of claim → justificaiton → uncertainty → conditions of rebuttal.

Item-test consistency. We examined how students' responses to each item were correlated with students' ability estimates measured by the entire test. Students who scored higher on an item should also score higher on the entire test since the items measure the same construct on the same scale. Figure 4 shows that students' average scientific argumentation ability estimates increased as scores increased in each of the four items in the Life item set. The same trend was found in all the other five item sets. We examined this relationship using item-test correlation values. Items with less than 0.3 item-test correlation values are considered to correlate little with the overall scale. Figure 5 shows item-test correlation values of claim, justification, uncertainty, and conditions of rebuttal items. All of the justfication and conditions of rebuttal items were highly correlated on the overall scientific argumentation scale. All but one uncertainty item also showed acceptable correlation values. However, except the claim item for the T2050 item set, claim items had less than 0.3 values for the item-total correlations, indicating student responses to the claim items were insufficient to predict their overall scientific argumentation ability.

---------- Insert Figure 4 and Figure 5 Here ----------

Wright Map. Figure 6 shows how items and students distributed on the scientific argumentation scale expressed in logit values from -4.0 to +4.0. On the left side, the distribution of students according to their scientific argumentation ability is shown. The higher on the scale, the more able students are on the scientific argumentation construct. On the right side, item thresholds of all scores in claim, justification, uncertainty, and conditions of rebuttal items are shown. An item threshold is defined as students with the matching ability would have a 50% chance of receiving a score *j* as compared to receiving a score *j – 1*. Since claim items were scored either 0 or 1, there was only one item threshold for each claim item. Justification items

scored from 0 to 4 had four item thresholds in each item. Uncertainty items had two item thresholds while conditions of rebuttal items had three. The higher the item threshold on the scale, the more difficult for students to receive that score on the item. The item threshold locations of scores within each of the justification, uncertainty, and conditions of rebuttal items indicate that the intial raw scores were merely on an ordinal scale. For example, item threshold values for the justification score 1 to the justification score 4 on the Ocean item set changed from -2.47, -0.12, 1.02, and 2.00, respectively, on the scientific argumentation scale. In this Rasch transformation from the raw scores to the Rasch estimates, only the order of the raw scores was preserved. The interval between the two adjacent raw scores (a score distance of 1) was not kept constant on the scientific argumentation scale, indicating that the amount of abilities assessed by the raw scores was not equally spaced. Nor were ratios between the two raw scores were kept constant.

---------- Insert Figure 6 Here ----------

For claim items, the scientific claim of the Ocean item set was most difficult and that of the Life item set was easiet. It was increasingly more difficult for students to receive a higher score in each justification item because the item thredhold values were located higher on the scale as justification scores increased. We created gray bands to locate the item threshold values for the same justification scores across justification items. The top end of each justification score band overlapped with the bottom end of the next justification score band. For uncertainty items, higher scientific argumentation abilities were needed for students to be certain about their arguments than to be neutral or to be uncertain. However, there was a large overlap between the certain and the neutral uncertainty score bands. For conditions of rebuttal, higher and higher scientific argumentation abilities were needed as students moved from citing personal reasons to discussing uncertainty within the context of investigations and to discussing uncertainty beyond

investigation. The three conditions of rebuttal score bands did not ovelap with one another.

The locations of these score bands across four types of items indicate that justification items covered the widest range of the scientific argumentation ability scale between -3.60 to +3.80. Conditions of rebuttal items covered the range of -1.35 to +3.10. The range covered by claim items was smaller than those covered by justification and conditions of rebuttal items but slightly larger than the range covered by the uncertainty items. Both uncertainty and claim items covered the middle ability range of the scientific argumentation scale.

The score band of making single warrants was located at the similar range to that of considering conditions of rebuttal within investigation. The score band of making two or more warrants was located at a similar range to that of conditions of rebuttal beyond investigation. These findings suggest that students who could make single warrants were more likely to consider conditions of rebuttal within investigation. Students who could make multiple warrants were more likely to consider conditions of rebuttal beyond investigation, indicating that students need to make multiple warrants based on multiple evidence pieces in order to consider limitations of the investigations imposed by current science, inquiry method, or other factors.

Reliability. The scientific argumentation scale shown in Figure 6 had the person separation reliability of 0.77 and the item separation reliability of 1.00. The item separation reliability was higher than the person separation reliability because the former was based on 837 students' responses to each item while the latter was based on 24 responses generated by a person. The Cronbach alpha value was 0.75, which was analogous to the person separation reliability.

Discussion

Recent science education reform movements put an immense emphasis on using scientific argumentation as a means to learn and teach science (Duschl et al., 2007) and

developing students' scientific argumentation ability over multiple school years (Smith, Wiser, Anderson, & Krajcik, 2006). Such continuous development of the scientific argumentation ability across courses, classes, and grade levels can be expedited through the employment of theoretically, psychologically, and psychometrically valid and reliable assessments. While most science education research on scientific argumentation has focused on the characterization and identification of scientific argument elements that appear in classroom discourses or written artifacts, a systematic large-scale assessment framework that can assist the longitudinal development of scientific argumentation has yet to be proposed. The most critical aspects of assessment development are related to defining an adequate construct, a type of cognition responsible for students' formulation of scientific arguments, developing items on the construct, and analyzing and interpreting assessment data on the construct (Pellegrino, Chudowsky, & Glaser, 2001).

In this study, we developed and validated a construct on scientific argumentation by combining theoretical and psychological interpretations of the existing work reported by others and psychometric interpretations of the findings reported in this work. At the theory level, we conceptualized the scientific argumentation construct by closely following Toulmin's argument structure (1958), widely adopted by the science education community. Toulmin's theory on argument structure identified six field-invariant elements in a rhetorical argument such as claim, data, warrant, backing, qualifier, and conditions of rebuttal. However, Toulmin's theory cannot specify how the analysis of these elements should be used to place students on a measurement scale. At the psychological level, we reviewed various analytic frameworks available in the literature (Clark et al., 2007; Samson & Clark, 2008) in order to develop scoring rubrics so that different levels of students' performances can be distinguished on claim, justification, uncertainty qualifier, and conditions of rebuttal. The main agenda of this paper therefore was

designing items to measure the scientific argumentation ability and applying psychometric analyses to student responses to the items.

Investigating construct validity of items can provide insights on cognitive theory used to design the items (Messick, 1989). Construct validity, in its simplest definition, concerns whether and how well the test is measuring what it is designed to measure. A test with construct validity can verify theoretically-identified relationships (Messick, 1994). Rasch analysis results in this study suggest that an underlying construct can account for students' responses to claim, justification, uncertainty qualifier, and conditions of rebuttal. The scientific argumentation construct appears to capture a complex ability that goes beyond knowing and understanding scientific content (Aufschnaiter, Erduran, & Osborne, 2008) because it also involves students' confidence in their arguments and recognition of limitations in the scientific data and the current knowledge. The involvement of complex cognitive activities in the scientific argumentation construct is predicted by Toulmin's theory (1958) and is verified in this study using item fit statistics. It is surprising to find out how well uncertainty items conform to the Rasch Partial Credit Model, as shown in item fit statistics, and how consistently they relate to the overall scientific argumentation ability, as shown in item-total correlations, even though uncertainty items may not be directly related to content.

Since we verified that the overall scientific argumentation ability consists of claim, justification, uncertainty qualifier, and conditions of rebuttal, we examined how the scientific argumentation construct should be established to represent items and students on the ability continuum. In *The Uses of Argument* (1958), Toulmin described the argument structure in the order of claim and data, warrant, qualifier, conditions of rebuttal, and backing. Some researchers asserted that a higher scientific argumentation ability can be seen in the number of argument elements identified in student discourse (Berland & McNeill, 2010) or written arguments (Bell &

Linn, 2000). Others suggested that being able to use counterarguments or conditions of rebuttal be an indication of higher scientific argumentation ability (Erduran et al., 2004; Kuhn, 2010). Kelly and Takako (2002) discovered that the task of ordering epistemic levels appeared to be difficult, if not impossible. In this study, we applied the Rasch Partial Credit Model to students' actual responses to see whether the order of the levels in the construct can be empirically established in terms of how much ability is required for students to achieve the levels. For this task, we examined average item difficulty values across claim, justification, uncertainty, and conditions of rebuttal items and item threshold values within and across item types.

Average item difficulty values across four scientific argumentation item types show the order of ability requirement increasing from uncertainty, to claim, to justification, and to conditions of rebuttal. The order between claim and justification reflects the general findings in the science argument and explanation literature that students have difficulty coordinating theory and evidence to justify their claim (Sandoval & Milwood, 2005) and that a higher ability is required to make links between scientifically-relevant ideas and salient evidence than to select or generate claims (Lee, Liu, & Linn, 2011; Liu, Lee, Hofstetter, & Linn, 2008). In addition, this study provides unique insights on the placement of uncertainty qualifiers and conditions of rebuttal on the scientific argumentation construct. Uncertainty acts as a vehicle to enable students to formulate scientific arguments. Since students' scientific argumentation depends upon content knowledge they have (Aufschnaiter et al., 2008; Millar & Driver, 1987), those who self-assess to have weak knowledge on the science topic formulate relatively weak scientific arguments as compared to those who have strong knowledge. This pattern is apparent in this study as more than two thirds of students' conditions of rebuttal citing personal reasons for their uncertainty rather than limitations of investigations. Therefore, student uncertainty in scientific arguments is characteristically different from scientists' uncertainty in the sense that the latter addresses

limitations of collective knowledge, understanding, and tools of science while the former primarily addresses self-concepts before it is transitioned to scientific uncertainty. Rarely are students in science classrooms asked to analyze the limitations of the school science presented to them (Lederman & O'Malley, 1990). It would not cross most students' minds that anything they are asked to do in science class could be uncertain. It can be beneficial to use uncertain science contexts such as climate change and life on other planets in the scientific argumentation instruction as a way to introduce tentativeness of current science knowledge and tools to students and teachers (McDonald, 2010).

Comparison of item threshold values for scores within each item confirms the increasing order of the levels used in scoring rubrics. Examination of item threshold locations show that the intervals between adjacent raw scores in each item are not equally spaced on the scientific argumentation scale. Thus, applying inferential statistics to students' initial scores on individual items is extremely problematic (Michell, 1999). The only safely allowed mathematical operations on variables on ordinal scales are the number of cases, mode, contingency correlations, median, and percentiles (Stevens, 1946). The sum of scores from different argumentation elements cannot represent students' overall scientific argumentation ability because that transformation does not give the isomorphic mapping between students' actual scientific argumentation abilities and their numerical estimates of their abilities (Krantz, Luce, Suppes, & Tversky, 1971). On the other hand, this challenge can be addressed to some extent by applying stochastic transcormations of raw scores like Rasch analyses (Karabatsos, 2001).

On the scientific argumentation scale, claim and uncertainty items cover the middle ability range while justification and conditions of rebuttal items cover the wider ability range. Justification items show the widest coverage, including very low and very high ability ranges. Conditions of rebuttal items extend the coverage towards the high ability range. This shows that

the scientific argumentation ability is mostly contributed by justification (Berland & Reiser, 2009) and conditions of rebuttal items, indicating that the assessment of scientific argumentation ability needs to focus on these two elements (Yeh, 2001). The Wright Map of the scientific argumentation scale also demonstrates that students are capable of considering the limitations of given investigations when prompted (Metz, 2004). However, for this to occur, students should be able to make links between theory and evidence, suggesting that scientific argumentation instruction can include not only justification but also conditions of rebuttal to improve students' scientific argumentation ability in a mutually reinforcing manner.

The assessment framework outlined and tested in this study can provide an effective means to assess and document development of students' scientific argumentation ability across science topics and over time. The item design and the scoring rubrics have general characteristics for manifestation depending upon science topics. The scientific argumentation scale shown in Figure 6 can afford multiple trajectories of student development within and across items such as (1) making a larger number of claims consistent with current science, (2) coordinating between theory and evidence more often, (3) becoming more certain about arguments they develop, and (4) considering limitations of the study related to investigations more often. Therefore, the overall scientific argumentation ability estimate can parsimoniously reflect students' performances on all required argument elements. Moreover, the Rasch ability estimates (Rasch, 1966) can be considered on an interval scale, making them suitable for the application of parametric inferential statistics (Bond & Fox, 2007).

The generalization of the findings in this study is limited due to the sampling of students and the limited number of items used in the scientific argumentation test. If we can include a larger number of higher ability students on the scientific argumentation scale, then item difficulty and threshold estimates of higher scoring levels can be made more accurately. The test we used

only considered climate change and the search for life on other planets. Inclusion of other science topics, in particular more defined science topics such as Newtonian Mechanics, may show different relationships among the four argument element item types from those we discovered in this study. Future research can examine a possibility of multi-dimensionality of the scientific argumentation construct, expand the assessment framework outlined in this study to other science topics and students at different grade and ability levels, use the framework along with interventions, and follow students' development over an extended period of time.

Conclusion

Informed by Toulmin (1958) and the current advances in scientific argumentation research, we conceptualized a psychometrically-valid and reliable scientific argumentation construct. To examine whether claim, justification, uncertainty qualifier, and conditions of rebuttal can be mapped onto an overall scientific argumentation construct, we applied the Rasch Partial Credit Model. The scientific argumentation scale established in this study greatly simplifies the assessment of students' scientific argumentation ability and provides a new analytic assessment model for studying learning progressions of scientific argumentation across science topics and disciplines. As the uncertain nature of science is an important epistemological belief about science, this study sheds light on the role students' interpretations of uncertainty play in formulating scientific arguments.

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant No. 0929774. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors gratefully acknowledge students and teachers who participated in this study.

References

American Association for the Advancement of Science (1993). Benchmarks for science literacy. New York: Oxford University Press.

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

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice-Hall.

Bell, P., & Linn, M. C. (2000). Scientific arguments as learning artifacts: designing for learning from the web with KIE. International Journal of Science Education, 22(8), 797-817.

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

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.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model. Mahwah, NJ: Lawrence Erlbaum Associates.

Bricker, L. A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. Science Education, 92, 473-493.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. Review of Educational Research, 63(1), 1-49.

Chinn, C., & Osborne, J. (2010). Students' questions and discursive interaction: Their impact on argumentation during collaborative group discussions in science. Journal of Research in Science Teaching, 47(7), 883-908.

Clark, D., Sampson, V., Weinberger, A., & Erkens, G. (2007). Analytic frameworks for assessing dialogic argumentation in online learning environments. Educational Psychology Review, 19, 343-374.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. Studies in Science Education, 38, 39-72.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: 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, 915-933.

Germann, P. J., & Aram, R. J. (1996). Student performances on the science process of recording data, analyzing data, drawing conclusions, and providing evidence. Journal of Research in Science Teaching, 33(7), 773-798.

Hogan, K., Nastasi, B. K., & Pressley, M. (2000). Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. Cognition and Instruction, 17(4), 379-432.

International Association for the Evaluation of Educational Achievement (IEA) (1995). TIMSS science items: Released set for population 1 (third and fourth grades). Chestnut Hill, MA: Boston College.

Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (1999). "Doing the lesson" or "doing science": Argument in high school genetics. Science Education, 84, 757-792.

Joint Committee for Guides in Metrology. (2008). International vocabulary of metrology-basic and general concepts and associated terms. Geneva, Switzerland: International Organization for Standardization.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. Journal of Applied Measurement, 2(4), 389-423.

Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university of oceanography students' use of evidence in writing. Science Education, 86, 314-342.

Kennedy, D. & Norman, C. (2005). 125 Questions: What don't we know? Science, sciencemag.org/sciext/125th/.

Koslowski, B. (1996). Theory and evidence: The development of scientific reasoning. Cambridge, MA: MIT Press.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). Foundations of measurement (Vol. I). Mineola, New York: Dover Publications, Inc.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. Science Education, 77(3), 319-337.

Kuhn, D. (2010). Teaching and learning science as argument. Science Education, 94(5), 810-824.

Kuhn, D., & Udell, W. (2003). The development of argument skills. Child Development, 74(5), 1245-1260.

Latour, B., & Woolgar, S. (1986). Laboratory life: The construction of scientific facts. Princeton, NJ: Princeton University Press.

Lawson, A. E. (2003). The nature and development of hypothetico-predictive argumentation with implications for science teaching. International Journal of Science Education, 25(11), 1387-1408.

Lederman, N. G., & O'Malley, M. (1990). Students' perceptions of tentativeness in science: Development, use, and sources of change. Science Education, 74(2), 225-239.

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. Applied Measurement in Education, 24(2), 115-136.

Lee, M.-H., Wu, Y.-T., & Tsai, C.-C. (2009). Research trends in science education from 2003 to 2007: A content analysis of publications in selected journals. International Journal of Science Education, 31(15), 1999-2020.

Lindley, D. (2007). Uncertainty: Einstein, Heisenberg, Bohr, and the struggle for the sole of science. New York: Anchor Books, A Division of Random House, Inc.

Linn, M. C., & Eylon, B.-S. (2006). Science education: Integrating views of learning and instruction. In P. A. Alexander & P. H. Winne (Eds.), Handbook of educational psychology (pp. 511-544). Mahwah, NJ: Lawrence Erlbaum Associates.

Liu, O. L., Lee, H.-S., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. Educational Assessment, 13(1), 1-23.

McDonald, C. V. (2010). The influence of explicit nature of science and argumentation instruction on preservice primary teachers' views of nature of science. Journal of Research in Science Teaching, 47(9), 1137-1164.

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.

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

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. Cognition and Instruction, 14(2), 139-178.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd Ed.) (pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. Cognition and Instruction, 22(2), 219-290.

Michell, J. (1999). Measurement in psychology. New York: Cambridge University Press.Millar, R., & Driver, R. (1987). Beyond processes. Studies in Science Education, 14, 33-62.

Mislevy, R. J., & Riconscente, M. M. (2005). Evidence-centered assessment design: Layers, structures, and terminology. Menlo Park, CA: SRI International.

National Research Council (1996). National science education standards. Washington, DC: National Academy Press.

Ohlsson, S. (1992). The cognitive skill of theory articulation: A neglected aspect of science education? Science and Education, 1(2), 181-192.

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.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing what students know. Washington, DC: National Academic Press.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology, 82(1), 33-40.

Rasch, G. (1966). Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press.

Sadler, T. D., & Fowler, S. R. (2006). A threshold model of content knowledge transfer for socioscientific argumentation. Science Education, 90, 986-1004.

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

Sampson, V., & Clark, D. (2009). The impact of collaboration on the outcomes of scientific argumentation. Science Education, 93, 448-484.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. The Journal of the Learning Sciences, 12(1), 5-51.

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.

Schwab, J. J. (1962). The teaching of science as enquiry. In J. J. Schwab & P. F. Brandwein (Eds.), The teaching of science (pp. 3-103). Cambridge: Harvard University Press.

Schwartz, B. B., Neuman, Y., Julia, G., & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activity. The Journal of the Learning Sciences, 12(2), 219-256.

Simosi, M. (2003). Using Toulmin's framework for the analysis of everyday argumentation: Some methodological considerations. Argumentation, 17, 185-202.

Smith, C., 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 atomic-molecular theory. Measurement: Interdisciplinary Research and Perspectives, 14(1&2), 1-98.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103(2684), 677-680.

Toulmin, S. (1958). The uses of argument. New York: Cambridge University Press.

van Eemeren, F. H., Grootendorst, R., Henkenmans, F. S., Blair, J. A., Johnson, R. H., Krabbe, E. C. W., et al. (1996). Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments. Mahwah, NJ: Lawrence Erlbaum Associates.

Walton, D., Reed, C., & Macagno, F. (2008). Argumentation schemes. New York: Cambridge University Press.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago, IL: Mesa Press.

Wu, M., Adams, R., & Wilson, M. (1997). ConQuest: Generalized item response modeling software. Melbourne, AU: ACER Press.

Yeh, S. S. (2001). Tests worth teaching to: Constructing state-mandated tests that emphasize critical thinking. Educational Researcher, 30(9), 12-17.

Yerrick, R. K. (2000). Lower track science students' argumentation and open inquiry instruction. Journal of Research in Science Teaching, 37(8), 807-838.

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

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. Journal of Research in Science Teaching, 39(1), 35-62

Table 1. A Construct Map for Scientific Argumentation Involving Claim, Justification, Uncertainty Qualifier, and Conditions of Rebuttal

	Description of the level	Toulmin (1958)	Student characteristics	Item design in this study
Level 1	Non-scientific			
Level 2	Scientific claim	Claim	Students think scientific claims can be made without support of evidence.	Claim
Level 3	Coordination between claim and evidence	Claim + data	Students recognize that adequate evidence is needed to support a claim.	Justification
Level 4	Reasoned coordination between claim and evidence	Claim + data + warrant/backing	Students can use theory or established knowledge to coordinate claim and evidence.	
Level 5	Modified, reasoned coordination between claim and evidence	Claim + data +warrant/backing + qualifier	Students recognize the uncertainty of claim given the strength of warrants.	Uncertainty
Level 6	Conditional, modified, reasoned coordination between claim and evidence	Claim + data +warrant/backing + qualifier + conditions of rebuttal	Students recognize conditions that the current claim may not be held by analyzing limitations related to measurements, current theory or model, and phenomena under investigation.	Conditions of rebuttal

Table 2. Conditions of Rebuttal Coding Rubric

Source of Uncertainty	Uncertainty source	Description of categories
No Information (Score 0)	• No response	• Did not respond to the related uncertainty item but answered the linked claim and explanation items.
	• Simple off-task responses	• Wrote "I do not know" or similar answers • Provided off-task answers
	• Restatement	• Restated the scientific claim made in the claim item. • Restated the uncertainty rating.
Personal (Score 1)	• Question	• Did/did not understand the question.
	• General knowledge/ability	• Did/did not possess general knowledge or ability necessary in solving the question. • Did/did not learn the topic (without mentioning the specific topic) • Can/cannot explain/estimate
	• Lack of specific knowledge/ability	• Did not know specific scientific knowledge needed in the item set.
	• Difficulty with data	• Did not make sense of data provided in the item..
	• Authority	• Mentioned teacher, textbook, and other authoritative sources.
Scientific-Within investigation (Score 2)	• Specific knowledge	• Referred to/elaborated a particular piece of scientific knowledge directly related to the item.
	• Specific data	• Referred to a particular piece of scientific data provided in the item.
Scientific-Beyond investigation (Score 3)	• Data/investigation	• Recognized the limitation of data provided in the item and suggested a need for additional data. • Mentioned that not all factors are considered.
	• Phenomenon	• Elaborated why the scientific phenomenon addressed in the item is uncertain.
	• Current science	• Mentioned that current scientific knowledge or data collection tools are limited to address the scientific phenomenon in the item.

Table 3. Distribution of Students' Responses across Scientific Argumentation Categories

N=837	Pinatubo	T2050	Ocean	Galaxy	Life	Spectra	All
(a) Claim							
• Scientific	58.3	26.1	21.8	70.3	84.7	36.4	49.6
• Non-scientific	41.5	70.7	77.5	28.9	12.4	53.9	47.5
• Missing	0.2	3.1	0.7	0.8	2.9	9.7	2.9
(b) Justification							
• Multiple warrants	0.7	2.1	3.0	1.6	5.5	1.1	2.3
• Single warrant	1.2	3.3	11.4	34.8	19.6	8.0	13.1
• Relevant evidence/knowledge	59.5	11.8	33.5	42.6	44.3	22.8	35.8
• No Evidence	35.7	57.6	44.7	16.1	23.8	35.0	35.5
• Blank/Offtask	2.6	22.1	6.7	4.1	3.9	22.4	10.3
• Missing	0.2	3.1	0.7	0.8	2.9	9.7	2.9
(c) Uncertainty							
• Certain	68.2	22.5	63.2	67.7	65.4	29.5	52.8
• Neutral	22.2	27.1	19.1	21.6	17.9	20.8	21.5
• Uncertain	8.5	47.2	16.6	9.1	10.4	37.9	21.6
• Missing	1.1	3.2	1.1	1.6	6.3	11.8	4.2
(d) Conditions of Rebuttal							
• Scientific-Beyond Investigation	2.7	1.8	1.4	5.0	4.2	1.2	2.7
• Scientific-Within Investigation	20.1	5.7	12.9	23.7	23.3	4.3	15.0
• Personal	33.3	63.6	44.9	35.2	24.3	52.0	42.2
• No information	42.8	25.7	39.7	34.5	41.9	30.7	35.9
• Missing	1.1	3.2	1.1	1.6	6.3	11.8	4.2

Table 4. Rasch Partial Credit Model Analysis Results

Items	Item difficulty	Infit mean square	Infit error	Outfit mean square	Outfit error
(a) Claims					
• Pinatubo	-0.57	1.03	0.07	1.03	0.07
• T2050	0.87	0.97	0.08	0.95	0.08
• Ocean	1.16	1.03	0.09	1.10	0.09
• Galaxy	-1.15	1.07	0.08	1.09	0.08
• Life	-2.24	0.98	0.11	0.93	0.11
• Spectra	0.20	1.00	0.08	1.00	0.08
mean item difficulty =	- 0.29				
(b) Justifications					
• Pinatubo	0.23	0.95	0.06	0.94	0.06
• T2050	0.65	0.93	0.05	0.91	0.05
• Ocean	0.10	0.94	0.04	0.94	0.04
• Galaxy	0.01	0.97	0.05	0.97	0.05
• Life	-0.30	0.95	0.04	0.95	0.04
• Spectra	0.73	0.94	0.04	0.93	0.04
mean item difficulty =	0.24				
(c) Uncertainty qualifiers					
• Pinatubo	-1.42	0.96	0.06	0.96	0.06
• T2050	0.24	1.08	0.05	1.13	0.05
• Ocean	-1.00	0.99	0.05	0.99	0.05
• Galaxy	-1.38	1.08	0.06	1.18	0.06
• Life	-1.29	0.97	0.06	0.97	0.06
• Spectra	-0.07	1.13	0.05	1.16	0.05
mean item difficulty =	- 0.82				
(d) Conditions of rebuttals					
• Pinatubo	0.89	1.04	0.05	1.05	0.05
• T2050	0.88	0.95	0.06	0.95	0.06
• Ocean	1.10	0.98	0.05	0.97	0.05
• Galaxy	0.57	1.05	0.04	1.04	0.04
• Life	0.72	1.04	0.04	1.06	0.04
• Spectra	1.07	0.97	0.06	0.98	0.06
mean item difficulty =	0.87				

Figure 1. Toulmin's argument structure (Toulmin, 1958, p.104)

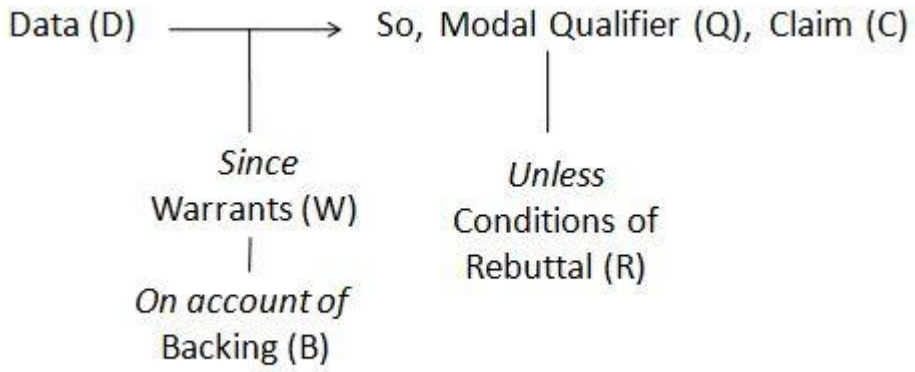


Figure 2.The Life item set is shown. The claim and justification items in the Life item set were modified from TIMSS (IEA, 1995, p.79).

LIFE ITEM SET

Jane and Mario were discussing what it might be like to live on other planets. Their science teacher gave them data about Earth and an imaginary planet, Athena. The table shows these data.

	Earth	Athena
Atmospheric Conditions	21% oxygen	10% oxygen
	0.03% carbon dioxide	80% carbon dioxide
	78% nitrogen	5% nitrogen
	ozone layer	no ozone layer
Distance from a Star Like the Sun	148,640,000 km	103,600,000 km
Rotation on Axis	1 day	200 days
Revolution Around Sun	365 ¼ days	200 days

Claim
Can life similar to Earth exist on Athena? Yes / No

Justification
Explain what might influence whether or not life can exist on Athena.

Uncertainty
How certain are you of your answer about life on Athena?
(1) Not at all certain
(2)
(3)
(4)
(5) very certain

Conditions of Rebuttal
Explain what influenced your uncertainty in the item above.

Figure 3. Scoring rubric for the justification item in the Life item set

Relevant knowledge or evidence:

- CO2 idea (C): Athena has much more CO2 than Earth
- Oxygen idea (O): Athena has less Oxygen than Earth
- Rev-Rot idea (R): Revolution/Rotation comparison (Athena's revolution and rotation periods are the same)
- Ozone idea (OZ): Athena does not have an ozone layer
- Difference idea (D): recognizing the difference between two locations.

Warrant links (coordinates between scientific knowledge and evidence to show why each piece of evidence is important)

- C link: more CO2 on Athena means hotter surface temperature than Earth
- O link: some earth-like life forms breathe oxygen Athena doesn't have enough oxygen
- R link: Athena's rotation and revolution periods are the same so one side of the planet is always facing the sun and therefore is hot while the other side is always dark and cold.
- OZ link: Harmful UV rays are blocked when there is an the ozone layer

(Score) Justification Levels	Criteria	Examples
(Score 0) Blank/ Off-task	• Did not write anything. • Wrote some text unrelated to the item.	• Blank answers • Because I think so. • Because Aliens live on Pluto and jupitor not Athena.
(Score 1) No knowledge/ evidence	• Restated the claim. • Elicited non-normative ideas. • Incorrectly mentioned the data. • Cited irrelevant data.	• Nothing matches Earth. • it looks normal • the details of Athena are relatively close to the details of EARTH • because carbon dioxide and nitrogen levels are high
(Score 2) Relevant knowledge/ evidence	• Mentioned that differences exist between two planets. • Listed data without mentioning how much difference exists. • Elicited one or more ideas listed above.	• There's not enough oxygen and too much CO2 • there is too much carbon and too little oxygen and there is no ozone layer because the environment is completely different. • all gases are different in level on Athena • the amount of oxygen, the distance to the star, the existence of an ozone layer
(Score 3) Warrant (=knowledge + evidence)	• Mentioned one of the warrant links listed above.	• there is no ozone layer which means if life was to form it would most likely get burnt up by the stars radiation.
(Score 4) 2 or more warrants	• Mentioned two or more of the warrant links above.	• The lower oxygen level would hurt any animal-like life. The increased level of carbon dioxide would increase the greenhouse effect, and it is much closer to the sun than Earth, so it would be much hotter, like Venus, and so life could not live there.

Figure 4. Average scientific argumentation ability estimates across scores in the Life item set

Figure 5. Item-test correlations across all items

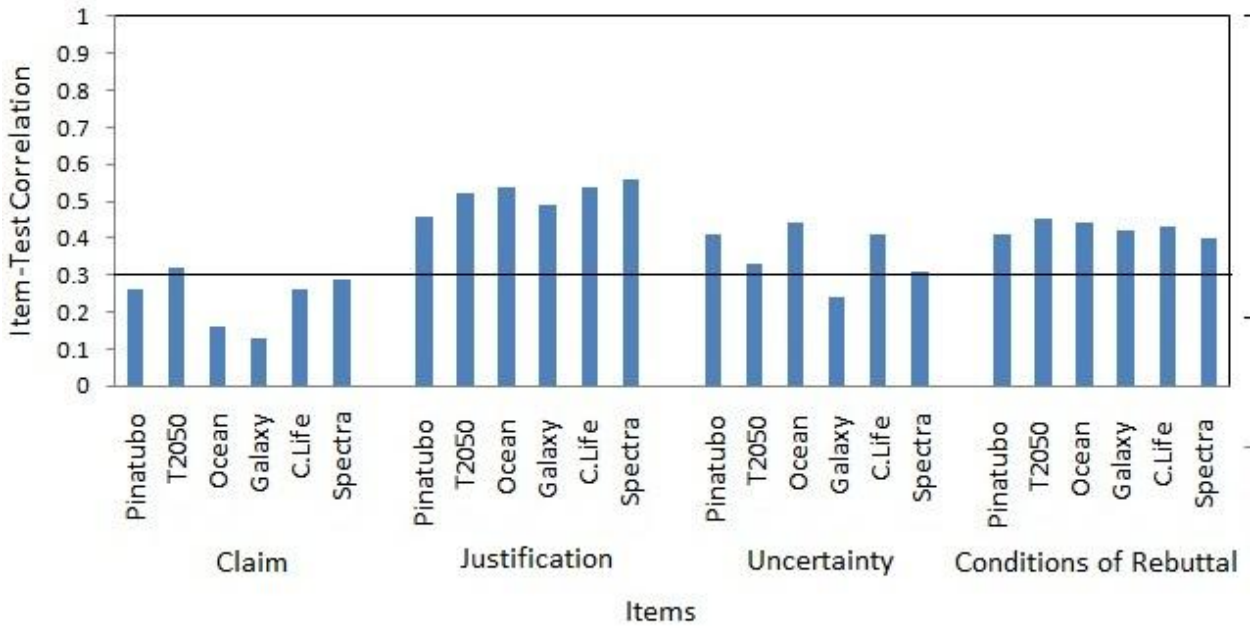
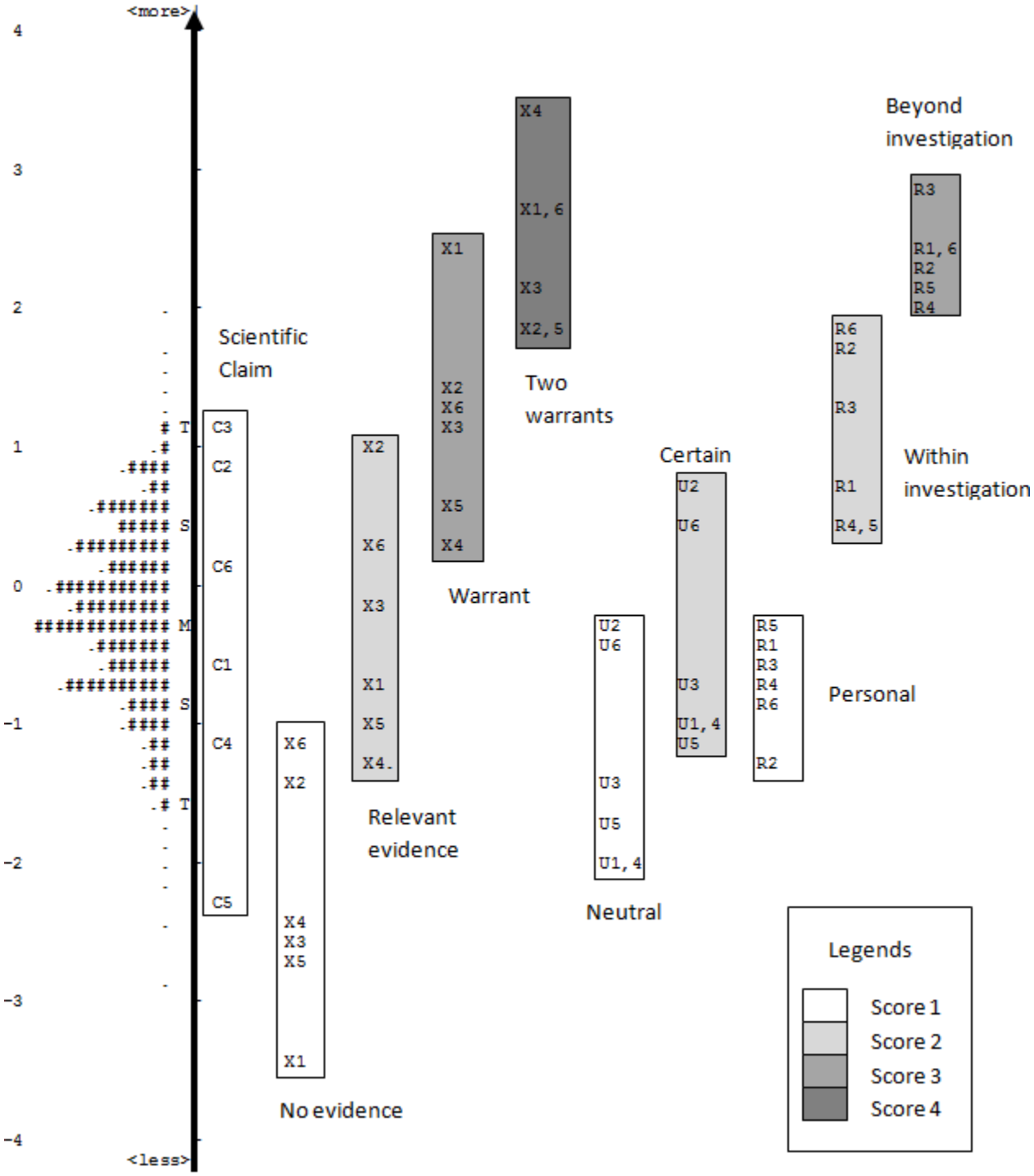


Figure 6. Wright Map



Note. “C” = Claim; “J” = Justification; “U” = Uncertainty; and “R” = Conditions of rebuttal; “1” Pinatubo Item Set; “2” T2050 Item Set; “3” Ocean Item Set; “4” Galaxy Item Set; “5” Life Item Set; “6” Spectra Item Set; “#” represents 7 students.