Characterizing Uncertainty Associated with Middle School Students' Scientific Arguments

Amy Pallant

The Concord Consortium

Hee-Sun Lee

Consultant to The Concord Consortium

University of California, Berkeley

Pallant, A., & Lee, H. –S. (2011). Characterizing uncertainty associated with middle school students' scientific arguments. Paper Presented at the annual meeting of the National Association for Research in Science Teaching, Orlando, FL.

Acknowledgements:
This material is based upon work supported by the National Science Foundation under the grant No. 0929774. Any opinions, findings, and 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 support and feedback from Sarah Pryputniewicz at the Concord Consortium and Dr. Ou Lydia Liu at the Educational Testing Service, Princeton, NJ.

Abstract

In this study, we investigated how students' claim, justification, uncertainty, and conditions of rebuttal contribute to the measurement of the overall scientific argumentation ability. We designed six sets of items, each of which consisted of claim, justification, uncertainty rating, and conditions of rebuttal items. These item sets addressed six investigations related to climate change and extraterrestrial life. We administered them to 956 students from 12 middle and high school teachers. We applied descriptive statistics and a Rasch Partial Credit Model analysis. Results of descriptive statistics show that students' difficulty in justifying their claims with scientifically valid warrants and in scientifically considering conditions of rebuttal that might undermine the strength of their argument. Rasch analysis results indicate that (1) all items can form a single scale, (2) students' scientific argumentation ability is represented in the order of uncertainty, claim, justification, and conditions of rebuttal, (3) justifications and conditions of rebuttal probe wider ranges of the scientific argumentation construct than claims and uncertainty ratings, (4) students who are able to 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.

Introduction

To make science learning authentic to actual science and meaningful to students' everyday lives, the use of scientific inquiry has been advocated (National Research Council, 1996). The process of scientific inquiry starts with a driving question, ensues with an investigation, and concludes with a claim and justification based on evidence collected from the investigation (Koslowski, 1996; Latour & Woolgar, 1985). Since the culminating step in scientific inquiry is communicating with others, 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; Lawson, 2003; McDonald, 2010; McNeill & Pimentel, 2010; Zembal-Saul, 2009; Zohar & Nemet, 2002). As a result, research on scientific argumentation has surged in the last decade (Lee, Wu, & Chai, 2008) with various frameworks proposed for analyzing rhetorical and dialogic arguments (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. Research has focused on analyzing claim, data as evidence, and warrant and backing as justification or reasoning. Studies typically define conditions of rebuttal as qualifiers and rebuttals as counterarguments and happened during dialogic discourse in the classroom, small groups, or online discussions. Studies on modal qualifiers such as uncertainty surrounding a claim given evidence in the context of scientific argumentation were rare, especially in rhetorical arguments by individuals.

Furthermore, there have been few attempts at examining whether and how students' responses to all these argument elements contribute to their overall scientific argumentation ability. In fact, most analytical frameworks tallied frequencies of occurrences. Occasionally, a sum of scores received across claim, data, and reasoning was used to represent students' overall scientific argumentation ability, even if scores in these argumentation categories may not be of the same type and cannot be assured of being on interval scales. Since scientific argumentation is advocated as an important science education learning outcome for all students, research on how to accumulate students' responses to these scientific argument

elements is needed so that students' overall scientific argumentation abilities can be documented, compared, and tracked over time at large scales.

In this study, we characterize secondary students' uncertainty and conditions of rebuttal, two less-frequently analyzed elements, in the context of rhetorical scientific arguments. The qualifier modifies the degree of the certainty of the claim based on evidence in an argument, and conditions of rebuttal show why the certain qualifier was chosen in the argument. Few scientific claims and justifications are made with absolute certainty by scientists due to incomplete or insensitive measurements, limitations in current theory or model, and phenomena under investigation (AAAS, 1993). We characterize scientific argumentation as a multi-level construct that can be measured by analyzing students' claims, justifications for the claims, uncertainty qualifiers, and conditions of rebuttal. We designed six sets of four items to elicit these elements in topics of climate change and life in space. The research questions of this study are:

- What type of claims, justifications, uncertainty ratings, and conditions of rebuttal do students provide in formulating rhetorical scientific arguments?
- How are students' claims, explanations, uncertainty, 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. We present results in the order of research questions listed above. Finally, we discuss implications of the results of this study for science teaching and for science education research.

Literature Review

Argument

Though argument and argumentation are interchangeably used in the literature without clear distinction, we will 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 (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. Arguments can be seen across disciplinary fields (field-invariant). Toulmin (1958) extracted six elements that are present in an argument:

- 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).

Rebuttals have been largely conceptualized as counterarguments in 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." According to Toulmin's terminology, current uses of qualifiers in the scientific education community are similar to conditions of rebuttal.

A graphical layout of an argument is shown in Figure 1. Toulmin (1958) pointed out that though these elements are field invariant, backing (B) provides "the criteria or sorts of ground required to justify" a claim (p. 36). Toulmin's field-invariant structure has been applied to analyze arguments across

disciplines. In case of scientific argument, the scientific knowledge base built upon the established and accepted scientific inquiry methods provide backing needed for the claim by means of warrants to a certain degree under conditions of rebuttals.

---------- Insert Figure 1 Hers ----------

Uncertainty and Conditions of Rebuttal in Scientific Argument

The most 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 answers. 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. Presence of uncertainty in a scientific argument modifies the strength of a claim made in the argument. Scientific uncertainty related to measurement, probability, phenomena, and status of current knowledge base can weaken the strength of the argument thus subject for rebuttal.

Measurement uncertainty. Measurement is a "process of experimentally obtaining one or more quantity values that can reasonably 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 because of the accuracy, precision, and resolution of the instrument.

Probability uncertainty. Scientific claims expressed in probability shows mathematical uncertainty. Probability describes the likelihood of a certain event to occur such as 60% chance of a shower tomorrow. Using probability distributions, all possible events are considered while none of the events are completely ruled out. Probability became extremely successful in addressing uncertainty especially in describing molecular, atomic, and subatomic phenomena.

Uncertain phenomena. Some scientific phenomena under investigation can be uncertain. The best example is Heisenberg's uncertainty principle where the position and the momentum of a particle cannot be measured accurately at the same time. This is the characteristic of the scientific phenomenon itself but the fault of the measurement method or the instrument. Moreover, 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. As complete understanding of any given scientific phenomenon is an almost impossible task, scientific claims cannot obtain absolute certitude due to unexamined components in the study.

Uncertainty due to current collective understanding. The knowledge, equipment, tools, and questions currently used by scientists limit claims and explanations scientists can provide. For example, on its 125th anniversary of publication, the *Science* magazine selected 125 questions that "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 is "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. In a study of elementary school students' inquiry-based investigations, Metz (2004) discover five spheres of students' uncertainty in how to produce a desired outcome, data, trend identified in the data, generalizability of the trend, and the theory that can explain the trend.

Developmental Trajectories in Scientific Argumentation

Most frameworks designed to analyze rhetorical or dialogic arguments put forward several ways to distinguish well-constructed from poorly-constructed arguments. Overall, an increasing competence has been identified in justifications, conditions of rebuttal, and counterarguments.

Justifications. Though students can engage in making arguments in everyday life and appear to be doing so quite naturally, they are not inclined to make arguments in science class. Often, students do not include their justifications for claims they put forward (Bell & Linn, 2000; Sandoval & Millwood, 2004). Justifications are often characterized as how students coordinate data or evidence with claim (Duschl & Osborne, 2002). Selecting salient evidence from available data for a particular claim is considered important (McNeill, Lizott, Krajcik, & Marx, 2006). Students' ability to justify is strongly

correlated with students' knowledge of science content relevant to the problem (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 for justifications focused on whether and how many scientifically-valid justifications are included. For instance, Clark and Sampson (2008) coded for the grounds students used in the order of claim only without grounds, data only, and 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), i.e. faults in reasoning shown 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 another argument while the latter not only is opposed to the original argument, but also has enough strength itself as an argument such that it overpowers the original argument. Conditions of rebuttal expressed in an argument are mainly to consider the first and the second types of rebuttals as they can weaken the strength of an argument to a certain extent. Students who realize claims 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 the third type of Walton et al. (2008)'s rebuttals as they are made as opposed to other claims with their own evidence and justifications. Analyses of counterarguments often occur in dialogic argument situations such as group or classroom discussions where opposite points of view can be elicited and debated. Arguments that consider potential counterarguments are more effective than without them (Erduran et al., 2004; Kuhn, 2010; Sadler &

Fowler, 2006). McNeill and Pimentel (2010) looked at claim, justification, and reasoning with counterarguments when assessing classroom argumentation.

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 have tried three methods. First, create a new set of categories by combining two or more argument element 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 higher scientific argumentation performances than the other four. Since this method creates categorical variables, only frequency comparisons are permitted.

Second, create a multi-level ordinal scale. 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 adds weak rebuttals to the second level while the fourth level adds one clearly identifiable rebuttal. The fifth level adds multiple clearly identifiable rebuttals. Osborne et al. (2004) used this ordinal scale to characterize to what level dialogic argument situation was able to reach in the classroom. 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 analysis of variance on this argumentation quality variable.

Third, create a total score by combining 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, Chin and Osborne (2010) used a composite score to find relationships between students' scientific argumentation ability and instructional practices. To score for each argument, counterargument, or rebuttal, Zohar and Nemet (2002) combined the number of justifications scored 0 to 2 and 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) combined 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 are observed for recognizing better responses within justifications, conditions of rebuttals, and counterarguments. However, the application of currently existing frameworks to students at large is limited because these argumentation variables have rarely been meaningfully accumulated to represent students' overall scientific argumentation ability.

Methods

In this section, we first define the scientific argumentation ability on a construct map (Wilson, 2004). We then describe our research as four-step assessment processes suggested by Mislevy and Ricoscente (2005): activity selection by assessment developers, activity presentation to students to collect data, evidence identification to collect salient information on target student performance, and evidence accumulation to amass student responses over multiple coding categories.

Rhetorical Scientific Argumentation Construct Map

Based on Toulmin's argument structure (1958), we conceptualized the rhetorical scientific argumentation construct consisting of six distinct levels. Table 1 shows these levels on a continuum in the order of increasing sophistication. Higher levels were assigned as students added 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 knowledge. In the third level, students make a claim based on data. In the fourth level, students make a claim based on evidence and elaborate their scientific reasoning related to how evidence leads to the claim. In the fifth level, students modify the strength of their scientific argument by recognizing limitations associated with measurement, current knowledge base or model, and phenomena. In the highest level, students can distinguish conditions that allow their modified scientific arguments to be held true from those that do not.

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

Instrument Design

We selected two science contexts, climate change and extraterrestrial life, from the 125 science problems a panel of scientists identified as "What We Don't Know" in *Science* (Kennedy & Norman, 2005). It was essential to select these current science topics to encourage students to elicit their uncertainty and conditions of rebuttal in their arguments. In cases where item contexts that scientifically correct answers are obvious, 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 CO2 levels between 125,000 years ago and 2000
- Ocean item set: predicting the trend of atmospheric CO2 level when ocean temperature increases

For the topic of extraterrestrial life, 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 based on 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 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 (Kelly, 1999; 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 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 level; the explanation items to the second and third levels; uncertainty items to the fourth level; the conditions of rebuttal items to the fifth level. Since the items were answered individually, 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 to be successful in the order of claim, explanation, uncertainty, conditions of rebuttal items.

Data Collection and Coding

We developed a test consisting of six item sets. The test was administered online to a total of 956 students taught by 12 teachers in six middle and high school schools in the Northeastern part of the United States. Among the students, 52% were female students; 90% spoke English as first language; 83% were middle school students; and 70% used computers regularly for homework.

Multiple-choice and short-answer claim items were dichotomously coded, "1" for scientific claim and "0" for non-scientific claim. Explanation items were coded based on whether scientifically relevant evidence was included and how well students reasoned with their included evidence. Figure 3 shows a scoring rubric for the explanation item in the Life item set. Explanations without science-related information were assigned to the no evidence category (score 1). Students can use as many as possible from the data provided in the Life item set such as differences between Athena and Earth in carbon dioxide, oxygen, revolution period, rotation period, and ozone layer. When justifications included relevant data but did not include how or why the data supported their claims, they were assigned to the relevant evidence category (score 2). Explanations that explained a link between the claim and data were assigned to the single warrant category (score 3). Students could be credited for making one out of the five possible links shown in Figure 3. When explanations provided two or more links between the claim and data, the two or more warrants category was assigned (score 4).

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

Student responses to uncertainty levels were scored into uncertain (score 0), neutral (score 1), and certain (score 2) categories. Student responses to conditions of rebuttal items were assigned into four categories as shown in Table 2. The first category represented blank, offtask responses, and restatements of claims or uncertainty ratings. The second category represented students' status of knowledge and ability related to the science topic addressed in the item. The third category dealt with scientific uncertainty involved in the outcome, knowledge, and data related to the investigation addressed in the item set. The fourth category represented scientific uncertainty that arise beyond the investigation.

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

Data Analysis

We used descriptive statistics to show what types of scientific claims, justifications, uncertainty levels, and conditions of rebuttal students in this study 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 shown below (PCM; Wright & Masters, 1982):

$$P_{nix}(\theta) = \frac{\exp[\sum_{j=0}^{x}(\theta_n - \delta_i - \tau_{ij})]}{\sum_{r=0}^{m_i}[\exp\sum_{j=0}^{r}(\theta_n - \delta_i - \tau_{ij})]} \quad (1)$$

where $P_{nix}(\theta)$ stands for the probability of student *n* scoring *x* on item *i* . θ stands for the student location on the knowledge integration construct in this study. δ_i refers to the item difficulty. τ_{ij} (*j* = 0, 1, ..*m*) is an additional step parameter associated with each score (*j)* for item *i*. We used the *Winstep* software (Linacre, 2010) to conduct the Rasch analysis. Using fit statistics, we first examined whether student responses to claim, justification, uncertainty, and conditions of rebuttal items could be interpreted on a single dimensional scale. We then examined overall item difficulties to determine how these four argumentation elements can be ordered according to the amount of ability required on the scientific argumentation scale. We also examined the Wright Map to compare the distributions of student abilities and item thresholds on the scientific argumentation scale. Item thresholds indicate how difficult for students to achieve a designated score within each item. On the Wright Map, we investigated the vertical ordering of the scores (i.e. whether higher abilities were needed to score higher on each item) and

the horizontal clustering of the scores (i.e. whether similar ability levels were needed for the same scores across the same item types and how required ability levels compare across claim, explanation, and uncertainty source items).

Results and Discussion

Student Response Distributions

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

Claims. Overall, 49.5% of the students' claims were scientific. The scientific claims related to T2050, Ocean, and spectra much less frequently occurred than the other three item sets. The lower scientific claim rates for T2050 and Ocean items may be related to students' difficulty with interpreting graphical representations that did not provide direct answers and with writing open-ended claims (note that the other four claims were multiple-choice claim items). For example, the T2050 item context showed the prehistoric global temperature graph and the level of atmospheric CO_2 concentration over the 125,000 year period. The ocean item showed the solubility of CO_2 in the ocean water 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.

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

Justifications. Overall, about half of the responses did not include any scientifically relevant evidence while slightly more than one third of the responses included salient evidence 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 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 more effectively formulated warrants with the Galaxy and Life items. Current science cannot provide definite claims related to whether life exists outside of Earth or

whether life exists based on a limited set of data. This indicates that students were more willingly engaged with scientific argumentation related to currently uncertain science.

Uncertainty. More than half of students' overall uncertainty responses 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 CO_2 solubility and atmospheric CO_2, leading the opposite claim to the scientific claim based on the graph.

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 investigations. In the example of the Life item set, students provided several issues that might undermine their arguments on the existence of extraterrestrial life form 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 a bacteria *[different life form might exist from what we know based on Earth life form]*" and "I'm sure there would be a way to make it work with advanced technology or some kind of manmade ozone layer *[advance in technology]*."

Rasch Scale for the Scientific Argumentation Construct.

Reliability. The person separation reliability was 0.74 while the item separation reliability was 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 traditional Cronbach alpha value was 0.75 which is analogous to the person separation reliability.

Item fit. Table 4 shows item fit statistics in mean square values. According to Bond and Fox

(2007), the acceptable range for item fit is between 0.70 and 1.30. There were no misfitting items based on infit statistics. Using outfit statistics, there was only one item outside of the acceptable range, the uncertainty rating item in the Galaxy item set with the outfit mean square value of 1.31. According to these results, all items can reasonably contribute to the measurement of the underlying overall scientific argumentation construct.

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

Figures 4 through 7 show how well students' actual responses fit the Rasch Partial Credit Model. In all figures, the x-axis indicates students' scientific argumentation abilities from low (-7.0) to high (7.0). The y-axis represents students' scores on the item. The Rasch Partial Credit Model represents a monotonically increasing relationship between student ability and student score on the item. That is, students are more likely to receive higher scores on the item as their underlying scientific argumentation abilities increase. Students' responses to justifications, uncertainty, and conditions of rebuttal items in the Life item set closely map onto the model lines. In the claim item, this monotonically increasing relationship holds except the very low ability students who picked the scientifically correct claim.

---------- Insert Figures 4 to 7 ----------

The order of item difficulty vs. the hypothesized order of scientific argumentation. We compared mean item difficulty values among claims, justifications, uncertainty qualifiers, and conditions of rebuttal. Table 4 shows that the easiest items were uncertainty qualifiers, followed by claims. The most difficult items were conditions of rebuttal. Justification items were placed between claims and conditions of rebuttal. These indicate that the order of scientific argumentation construct according to the required amount of ability should be revised to uncertainty → claim → justification → conditions of rebuttal.

The scientific argumentation scale. Since all items show acceptable fit to the Rasch Partial Credit Model, we establish a measurement scale for the overall scientific argumentation construct. Figure 8 shows how items and students distribute over this scientific argumentation scale. The logit scale in Figure 8 ranges from -4.0 to +4.0. On the left side, the student distribution according to their scientific argumentation ability is shown. On the right side, item thresholds of all scores in claim, justification, uncertainty, and conditions of rebuttal items are shown. The higher on the scale, the more able students are on the scientific argumentation construct. The higher on the scale, the more difficult for students to

receive the corresponding score on the item.

---------- Insert Figure 8 ----------

For claims, the scientific claim of the Ocean item set was most difficult to make and that of the Life item set was easiet to make. For justifications, it became increasingly more difficult for students to receive higher scores in their justifications. The order of justification difficulty shows no evidence → relevant evidence → single warrant → two or more warrants. The top end of each justification score band overlapped with the bottom end of the next justification score band. For uncertainty, higher scientific argumentation abilities were needed for students to be certain about their arguments than to be neural or uncertain. However, there was a large overlap between the certain and neutral uncertainty score bands. For conditions of rebuttal, higher and higher scientific argumentation abilities were required as students move 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 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 abiity 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, and the band of making two or more warrants was located at the similar range to that of conditions of rebuttal beyond investigation. These findings suggests that students who can make warrants are more likely to consider conditions of rebuttal within investigations. Students who can make two or more warrants are more likely to consider conditions of rebuttal beyond investigations, 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.

Conclusion

Interests in scientific argumentation have been increasing among science educators and researchers (Lee et al., 2008) and broadened as a means to promote, analyze, and assess student understanding in science through inquiry. In this study, we refined Toulmin's theory on the argument structure by establishing the underlying scientific argumentation construct in the order of increasing ability requirement. We also compared claims, justifications, uncertainty qualifiers, and conditions of rebuttal on the scientific argumentation scale resulting from the Rasch Partial Credit Model analysis. This uni-dimensional scientific argumentation scale can be used to effectively summarize student data collected from various elements of the argument structure. The scientific argumentation scale greatly simplifies large scale research on scientific argumentation 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 shed light on the role students' uncertainty plays in formulating scientific arguments.

References

AAAS (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

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.

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.

Chin, 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.

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.

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.

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.

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.

Linacre, J. M. (2010). Winsteps (Version 3.70.0)) [Computer Software]. Chicago: Winstep.com

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. (2009). Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena. *Science Education, 93*(2), 233-268.

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.

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.

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

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.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educatioanl Researcher, 18*(1), 16-25.

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. (2009). The impact of collaboration on the outcomes of scientific argumentation. *Science Education, 93*, 448-484.

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.

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.

Sismondo, S. (2004). *An introduction to science and technology studies*. Malden, MA: Blackwell.

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

Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning* (2nd ed.). New York: Macmillan.

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

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.

Zeidler, D. L., Walker, K. A., Ackett, W. A., & Simmons, M. L. (2002). Tangled up in views: Beliefs in the nature of science and responses to socioscientific dilemmas. *Science Education, 86*, 343-367.

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 uncertainty

	Description of the level	Toulmin (1958)	Student characteristics	Item design in this study
Level 0	Non-scientific			
Level 1	Scientific claim	Claim	Students think scientific claims can be made without support of evidence.	Claim
Level 2	Coordination between claim and evidence	Claim + data	Students recognize that adequate evidence is needed to support a claim.	Justification
Level 3	Reasoned coordination between claim and evidence	Claim + data + warrant/backing	Students can use theory or established knowledge to coordinate claim and evidence.	
Level 4	Modified, reasoned coordination between claim and evidence	Claim + data +warrant/backing + qualifier	Students recognize the uncertainty of claim by analyzing limitations related to measurements, current theory or model, and phenomena under investigation.	Uncertainty
Level 5	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.	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.
	• 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							
• 2 (Evidence + Warrant)	0.7	2.1	3.0	1.6	5.5	1.1	2.3
• Evidence + Warrant	1.2	3.3	11.4	34.8	19.6	8.0	13.1
• Relevant Evidence	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					
• C.Pinatubo	-0.59	1.04	0.07	1.04	0.07
• C.T2050	0.89	0.97	0.08	0.94	0.08
• C.Ocean	1.18	1.04	0.09	1.17	0.09
• C.Galaxy	-1.19	1.11	0.08	1.15	0.08
• C.Life	-2.31	0.98	0.11	0.93	0.11
• C.Spectra	0.19	1.00	0.08	1.01	0.08
mean item difficulty =	- 0.31				
(b) Justifications					
• J.Pinatubo	0.24	0.92	0.06	0.91	0.06
• J.T2050	0.72	0.91	0.05	0.87	0.05
• J.Ocean	0.12	0.91	0.04	0.91	0.04
• J.Galaxy	-0.02	0.95	0.05	0.96	0.05
• J.Life	-0.32	0.93	0.04	0.93	0.04
• J.Spectra	0.77	0.89	0.04	0.90	0.04
mean item difficulty =	0.25				
(c) Uncertainty qualifiers					
• U.Pinatubo	-1.5	0.96	0.06	0.97	0.06
• U.T2050	0.26	1.10	0.05	1.21	0.05
• U.Ocean	-1.06	0.98	0.05	1.01	0.05
• U.Galaxy	-1.46	1.12	0.06	1.31	0.06
• U.Life	-1.36	0.96	0.06	0.97	0.06
• U.Spectra	-0.07	1.16	0.05	1.24	0.05
mean item difficulty =	- 0.87				
(d) Conditions of rebuttals					
• R.Pinatubo	0.94	1.04	0.05	1.06	0.05
• R.T2050	0.93	0.92	0.06	0.92	0.06
• R.Ocean	1.16	0.96	0.05	0.95	0.05
• R.Galaxy	0.6	1.05	0.04	1.05	0.04
• R.Life	0.76	1.05	0.04	1.09	0.04
• R.Spectra	1.13	0.97	0.06	0.98	0.06
mean item difficulty =	0.92				

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

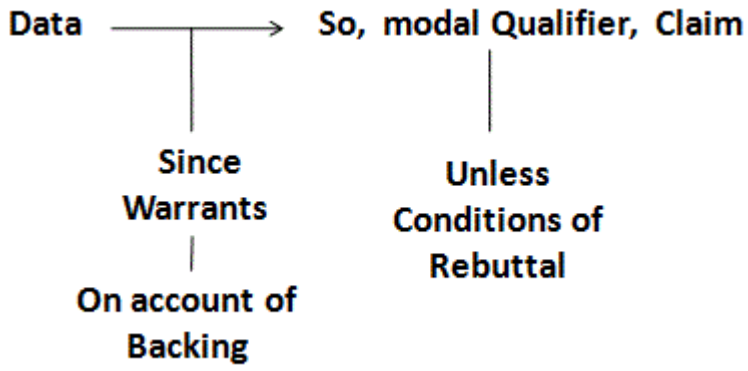


Figure 2. An explanation-uncertainty item set (Italics were added to indicate the item composition). The item was modified from TIMSS (IEA, 1995).

Jane and Mario were discussing what it might be like to live on other planets. Their scienc teacher gave them data about Earth and an imaginary planet, Athena. The table shows thes 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 1/4 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 question #19.

Figure 3. Explanation coding based on the knowledge integration scoring rubric
Relevant evidence:

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

Evidence + warrant links (explains 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
- R link: Athena's rotation and revolution periods are the same since 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 by 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 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 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) Evidence + Warrant	• Mentioned one of the evidence + 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 (Evidence + Warrant)	• Mentioned two or more of the evidence + 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. Item Characteristic Curve: Claim Item in the Life Item Set

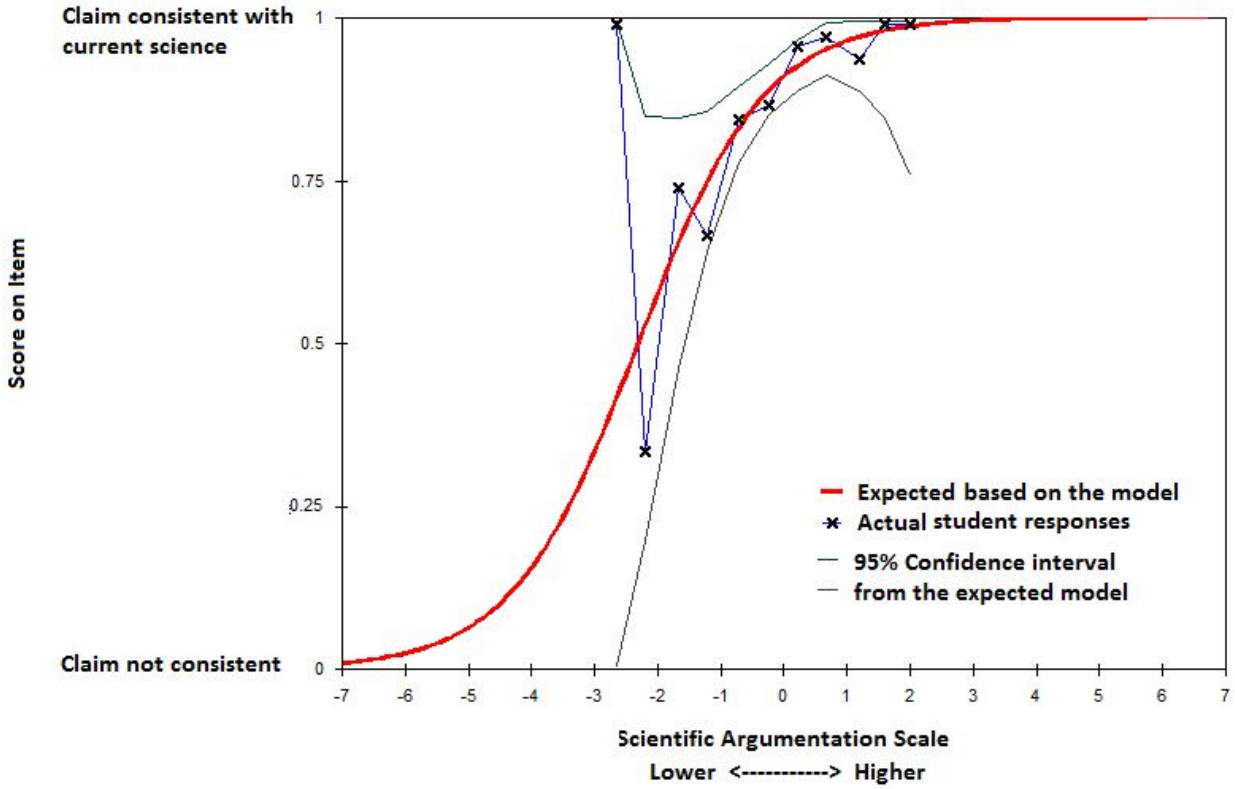


Figure 5. Item Characteristic Curve: Justification Item in the Life Item Set

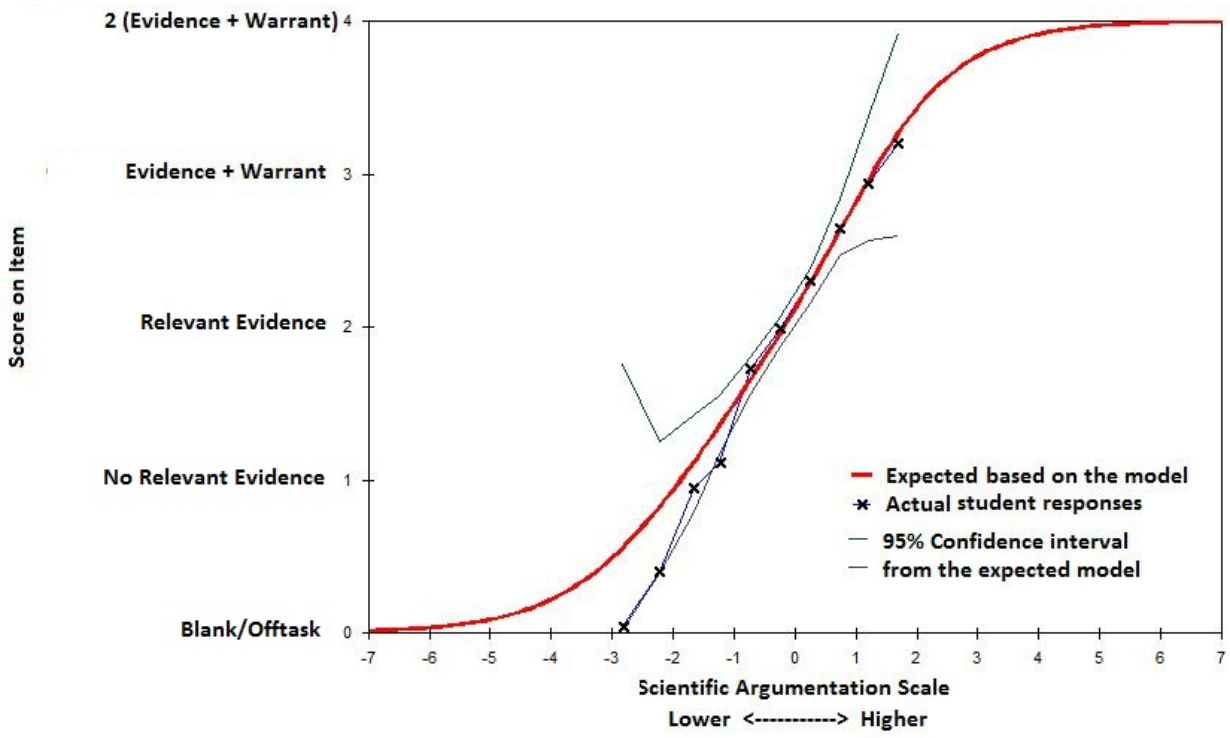


Figure 6. Item Characteristic Curve: Uncertainty Item in the Life Item Set

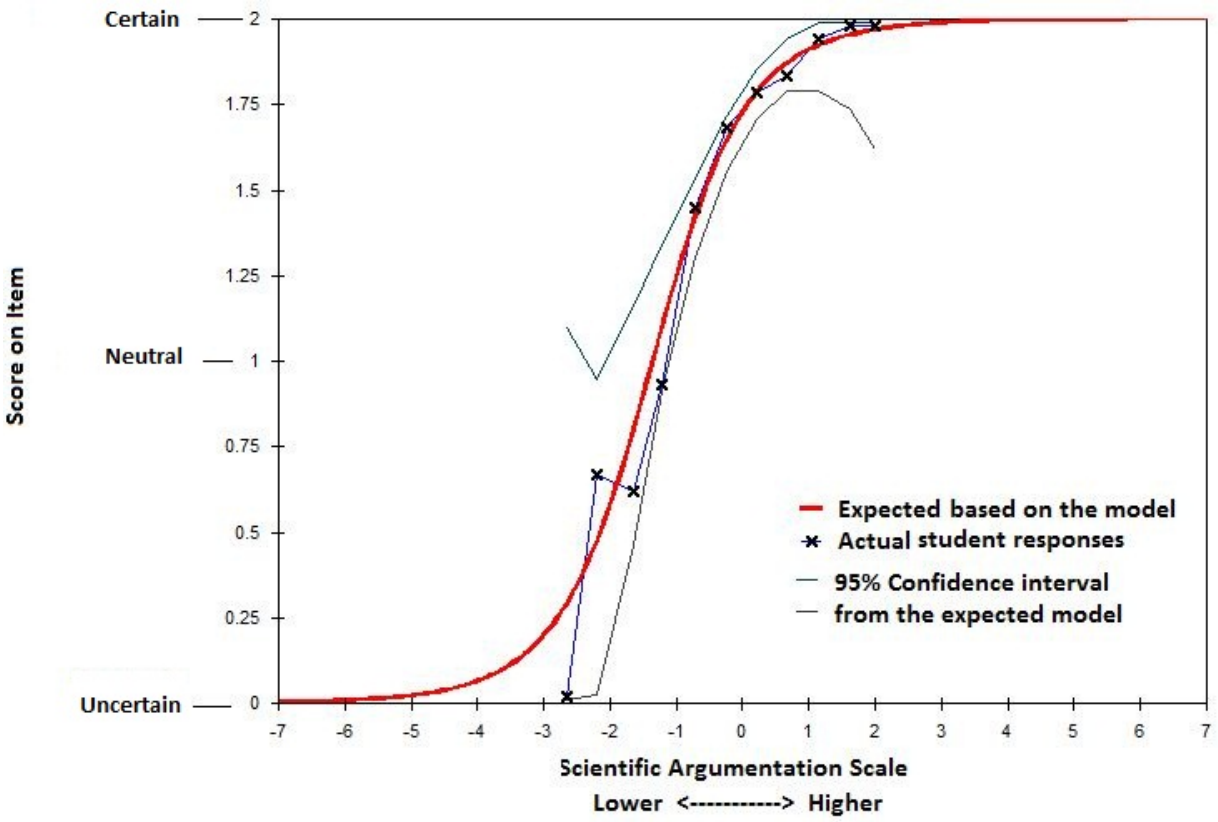


Figure 7. Item Characteristic Curve: Conditions of Rebuttal Item in the Life Item Set

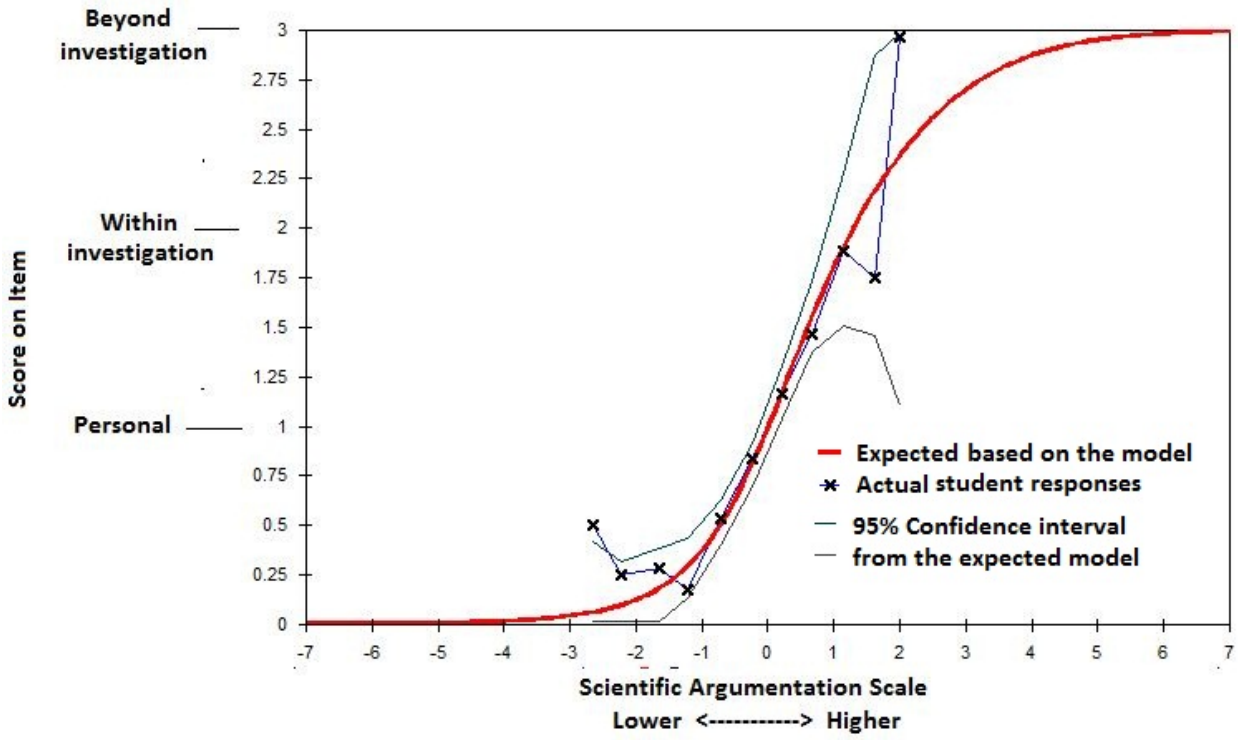
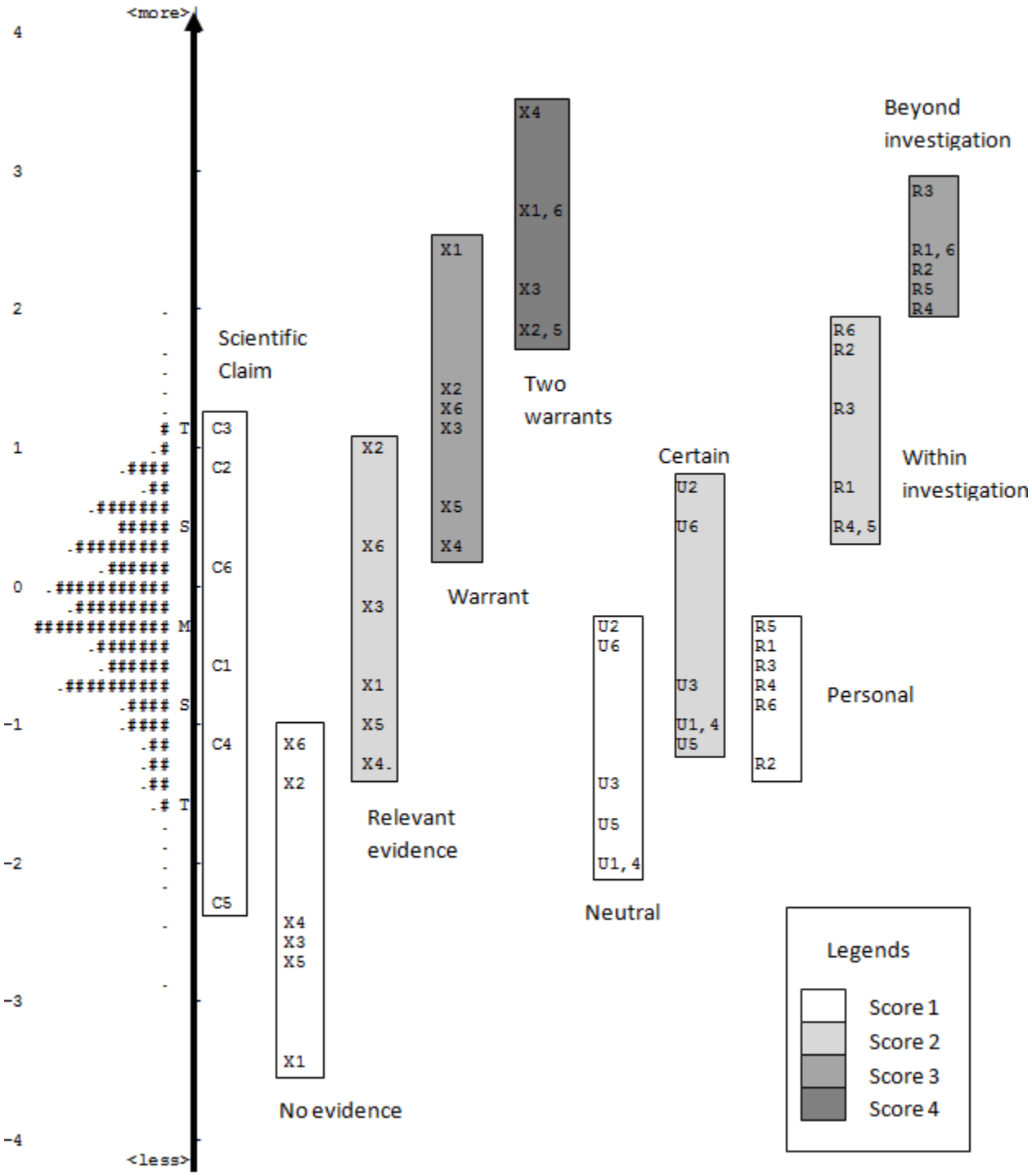


Figure 8. Wright Map.



Note. “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.