Chapter 12
The Development and Validation of the Assessment of Scientific Argumentation in the Classroom (ASAC) Observation Protocol: A Tool for Evaluating How Students Participate in Scientific Argumentation

Victor Sampson, Patrick J. Enderle, and Joi Phelps Walker

Introduction

Argumentation that is scientific in nature is often described as a form of "logical discourse whose goal is to tease out the relationship between ideas and evidence" (Duschl, Schweingruber & Shouse, 2007, p. 33) or a knowledge building and validating practice in which individuals propose, support, critique, and refine ideas in an effort to make sense of the natural world (Driver, Newton & Osborne, 2000; Kuhn, 1993; Sampson & Clark, 2011). Scientific argumentation, as a result, plays a central role in the development, evaluation, and validation of scientific knowledge and is viewed by many an important practice that makes science different from other ways of knowing (Driver et al., 2000; Duschl & Osborne, 2002). Yet, few students are given a opportunity to develop the knowledge or skills needed to participate in scientific argumentation or to learn how scientific argumentation differs from other forms of argumentation by time they graduate from high school (Duschl, et al., 2007; National Research Council, 2005, 2008) or as part of their undergraduate science education (National Research Council, 1999; National Science Foundation, 1996).

In response to this issue, several new instructional approaches and curricula have been developed over the last decade to give students more opportunities to acquire the understandings and abilities needed to participate in scientific argumentation. A continual challenge associated with this type of research, however, is the difficulties associated with documenting the nature or quality of the scientific

An earlier version of this chapter, entitled "Assessment of argumentation: An observation protocol," by P. Enderle, J. Walker, C. Dorgan, and V. Sampson was presented at the 2010 International Conference of the National Association for Research in Science Teaching.

V. Sampson (✉)
Florida State University, Tallahassee, FL, USA
e-mail: vsampson@fsu.edu

M.S. Khine (ed.), *Perspectives on Scientific Argumentation*,
DOI 10.1007/978-94-007-2470-9_12,

argumentation that takes place between students inside the classroom and tracking how students' ability to participate in scientific argumentation changes over time. Many researchers, for example, assess argumentation quality by first video or audio recording students as they engage in this complex practice, then they transcribe the discourse, and finally code or score it using a framework that focuses on the nature and function of the contributions to the conversation (Duschl, 2007; Erduran, 2007; Erduran, Simon & Osborne, 2004; Kelly, Druker & Chen, 1998; Kuhn & Reiser, 2005; Kuhn & Udell, 2003; Osborne, Erduran & Simon, 2004; Sampson & Clark, 2011). Yet, along with the various affordances that are associated with this type of approach for assessing the quality of argumentation, there are numerous constraints that stem from video taping, transcribing, and then "coding and counting" (Suthers, 2006).

First, this type of analysis is often time-consuming and expensive. Researchers, as a result, tend to study small samples of students or focus on a specific context. Second, the various aspects of a verbal argument are often difficult to identify during a discussion, which, in turn, tends to have an adverse effect on reliability (see Duschl, 2007; Erduran, 2007; Erduran, Simon et al., 2004). Another barrier to this type of approach is the nonlinear nature of scientific argumentation, which often makes it difficult for researchers to follow a line or thought through an episode of a multi-voiced argumentation or to define the boundaries of a unit of analysis. An analysis that focuses on the nature or functions of contributions to a discussion (e.g., the number of times the students support their claims or challenge the ideas of others, etc.) also limits what researchers are able to measure and forces them to disregard aspects of scientific argumentation that might be important or informative. For example, there are few studies that have examined the reasoning students employ during an episode of argumentation, the criteria they use to assess the merits of an idea, and how students interact with each other and the available materials as a way to assess quality. An assessment of argumentation quality that relies on a tabulation of the nature and function of contributions, therefore, is often limited in scope and privileges certain elements of argumentation at the expense of others.

The field therefore needs to develop new instruments that researchers can use to capture and score an episode of argumentation in a more holistic fashion, including nonverbal social interactions, and will result in a more comprehensive assessment of the overall quality of an event. Such an instrument also needs to be able to provide researchers with a reliable criterion-referenced measure of students' competency. This type of instrument is needed, as Driver et al. (2000) suggests, "to inform educational interventions designed to improve the quality of argumentation" and "to inform teachers about what to look for" (p. 295).

In this chapter, we will present a new instrument that researchers can use to measure that nature and quality of scientific argumentation. This instrument, which we call the *Assessment of Scientific Argumentation inside the Classroom* (ASAC) observation protocol, is intended to provide a criterion-referenced tool that targets the conceptual, cognitive, epistemic, and social aspects of scientific argumentation. This tool can be used to assess the nature and quality of argumentation that occurs between students inside the science classroom, to examine how students' participation in scientific argumentation changes over time in response to an intervention, or to compare the impact of different interventions on the way students participate in scientific argumentation. In the paragraphs that follow, we will first describe the method we used to develop and validate this new instrument. We will then conclude the chapter with a presentation and discussion of our findings, the limitations of work, and our recommendations about how and when to use the ASAC.

Method

Framework Used to Establish the Validity and Reliability of the ASAC

There is a large body of literature concerned with issues related to the validity and reliability of the assessment instruments that are used in educational research. In this methodological-focused literature, the legitimacy of drawing a conclusion about the knowledge or skills of people from scores on an assessment instrument is of upmost importance and shapes how new instruments are developed and validated. Many instrument developers, for example, use experts to determine if a new instrument measures what it purports to measure and estimates of internal consistency, such as Cronbach Alpha or KR-20, to evaluate the reliability of the instrument as part of the development and validation process (Burns, 1994). Although both are important, these two approaches are not sufficient. Trochim (1999), for example, suggests researchers need to focus on multiple properties of an instrument, such as the construct and criterion-related validity of an instrument, as well as its reliability in order to determine if an assessment actually measures what it is intended to measure. We therefore developed the methodological framework provided in Fig. 12.1 to guide the development and the initial validation of the ASAC.

In this framework, an instrument is deemed to possess good construct validity (i.e., the translation of a construct into an operationalization) if the theoretical construct is well defined, based on the available literature, and measures only the

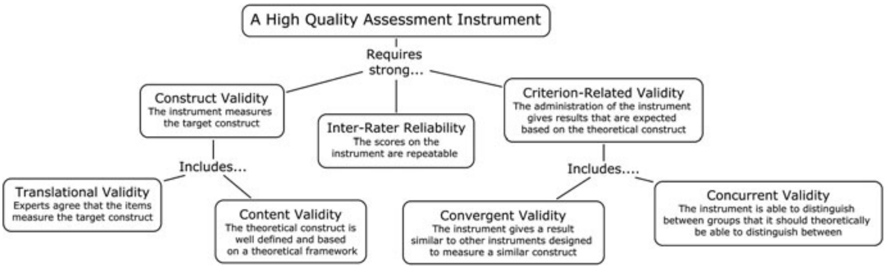
A High Quality Assessment Instrument
Requires strong...
Construct Validity
The instrument measures the target construct
Inter-Rater Reliability
The scores on the instrument are repeatable
Criterion-Related Validity
The administration of the instrument gives results that are expected based on the theoretical construct
Includes...
Translational Validity
Experts agree that the items measure the target construct
Content Validity
The theoretical construct is well defined and based on a theoretical framework
Includes....
Convergent Validity
The instrument gives a result similar to other instruments designed to measure a similar construct
Concurrent Validity
The instrument is able to distinguish between groups that it should theoretically be able to distinguish between

Fig. 12.1 The framework used to guide the development and initial validation of the ASAC

targeted construct (content validity). The items included in the instrument also need to be good translations of the construct based on expert opinion (translational validity). In addition to construct validity, an instrument must also have strong criterion-related validity in order to be considered credible and of high quality. Criterion-related validity considers the conclusions that can be drawn based on data generated by using the instrument. An instrument is deemed to possess strong criterion-related validity if it results in scores that are expected based on the theoretical construct (Trochim, 1999). An instrument with strong criterion-related validity, for example, should give results similar to another method that measures the same or a similar construct (convergent validity) and should be able to distinguish between groups or individuals that it is expected to distinguish between (concurrent validity). Assessment instruments, especially observation protocols, also need to have strong inter-rater reliability in order to generate scores that are consistent and repeatable.

Development of the ASAC

The method that we used to develop the ASAC began with a search of the literature in order to define the construct to be assessed and continued through item pool preparation, item refinement, and selection based upon expert review to ensure construct validity. The development then concluded with an evaluation of the instrument's criterion validity and reliability (Borg & Gall, 1989; Nunally, 1970; Rubba & Anderson, 1978). The seven-step process described below, which is based on recommendations outlined in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999), was used to create a high-quality assessment instrument that is both valid and reliable.

Step 1: Define the construct to be measured. A clear definition of the construct that an assessment is intended to measure is needed in order to guide the development of an instrument. The construct also needs to be well defined in order to evaluate the content validity of the instrument and to determine how well an instrument measures the construct of interest. Therefore, in order to guide our work, we adopted a view of argumentation as a process where "different perspectives are being examined and the purpose is to reach agreement on acceptable claims or course of actions" (Driver et al., 2000, p. 291). This view of argumentation stresses collaboration over competition and suggests that activities that promote argumentation can provide a context where individuals are able to use each other's ideas to construct and negotiate a shared understanding of a particular phenomenon in light of past experiences and new information (Abell, Anderson & Chezem, 2000; Andriessen, Baker & Suthers, 2003; Boulter & Gilbert, 1995; deVries, Lund & Baker, 2002; Veerman, 2003). In other words, argumentation is a social and collaborative process that groups of individuals engage in "to solve problems and advance knowledge" (Duschl and Osborne, 2002, p. 41). An important distinction in this definition stems from the focus on the *process* involved in argumentation, a less researched phenomena, as opposed to arguments, the *product* of such activities that has received more attention. Given this theoretical perspective, we chose to define the construct of *scientific argumentation* as a social and collaborative process of proposing, supporting, evaluating, and refining ideas in an effort to make sense of a complex or ill-defined problem or to advance knowledge in a manner that is consistent with conceptual structures, cognitive processes, epistemological commitments, and the social norms of science (see Driver et al., 2000; Duschl, 2008; Kuhn, 1993; Sampson & Clark, 2009).

Step 2: Development of the instrument specifications (content coverage and item format). Our goal at this stage of the development process was to ensure that the instrument would measure each aspect of the target construct as defined by our theoretical framework. To accomplish this goal, we decided to focus on three aspects of scientific argumentation, which according to Duschl (2008), students need to develop in order to be able to participate in this complex practice (p. 277). First, an individual must be able to use important conceptual structures (e.g., scientific theories, models, and laws or unifying concepts) and cognitive processes valued in science when reasoning about a topic or a problem. Second, an individual must know and use the epistemic frameworks that characterize science to develop and evaluate claims. Third, and perhaps most importantly, individuals that are able to engage in scientific argumentation must understand and be able to participate in the social processes that shape how knowledge is communicated, represented, argued, and debated within scientific community. We therefore decided to develop a protocol that was divided into four sections to represent these various aspects, each with a distinct focus for assessing an episode of scientific argumentation (conceptual, cognitive, epistemological, and social).

At this stage in the development process, we also decided to include items in the protocol that are observable during an episode of argumentation regardless of context or topic of discussion. We felt that this was important decision to make at the onset of the project because some elements of scientific argumentation, although important, cannot be measured easily through direct observation (e.g., it only take place within the mind of an individual, it only occurs when someone produces a formal written argument, or it tends to only occur under certain circumstances). Finally, and perhaps most importantly, we decided to craft the items so researchers can use them to rate an element of each aspect of scientific argumentation on a scale. We decided to use a rating scale rather than a simple dichotomous option (yes/no) so researchers will be able to document the prevalence of each element of the four aspects of scientific argumentation (e.g., not at all, often, etc.), which in turn, will allow for a greater distribution of scores.

Step 3: Development of the initial pool of items. We generated an initial pool of 29 items based on important notions and issues raised in the argumentation literature. Each item contains a stem sentence describing a critical element of an aspect of scientific argumentation and a detailed description offering insight into the aim of the item. These initial stems were written based on recommendations made by Edwards (1957) to reduce item error due to ambiguity. A Likert-style scale, which ranges from 0 to 3, was also included for ranking the element based on the presence and prevalence of the observable actions described in the stem sentence (0—not at all, 1—once or twice, 2—a few times, 3—often). A few items focused on undesirable actions in regards to quality scientific argumentation and, as such, the scaling of these particular items is reversed for scoring.

Step 4: Initial expert review of the item pool. We then conducted an evaluation of the content and translational validity of the initial pool of items. To complete this evaluation, we asked a group of experts to review the items using an online survey instrument. We identified 18 experts based on their significant contributions to argumentation research and the relevant literature. We then sent the group of experts an email to explain the objective of this project, a request for their service, and a link to the online survey. We asked the reviewers to rank whether each item was an important aspect of scientific argumentation and if it should be included in the protocol on a scale of 1–5 (five being the highest). We also asked the reviewers whether the description for each item was appropriate and offer suggestions about how an item or the description of the item should be revised. The reviewers' identities were kept anonymous to allow for the utmost candor in their responses.

The online survey was kept active for a period of 2 months following the initial email that was sent in order to solicit the services of the experts. In total, eight thoughtful responses were received and used to make adjustments to the observation protocol and the initial item pool. Items that received an average ranking of 4 or higher (i.e., agree to strongly agree) were kept, the items with an average ranking between 3 and 4 (neutral to agree) were revised or combined with other items based on the reviewer's comments, and the items that had an average ranking of between 1 and 3 (i.e., strongly disagree to neutral) were discarded. This process resulted in the elimination of a total of eight items from the initial collection.

Step 5: The first field test of the instrument. The next phase of development process involved the authors attempting to use the protocol to assess the quality of several video-recorded episodes of argumentation. Our objective at this stage of the development process was to ensure that a rater could observe the element of argumentation targeted by an item and to ensure that all the items were clear and the accompanying descriptions were detailed enough to produce reliable scores across multiple raters. To accomplish this task, we viewed several videos of students engaged in a task that was designed to promote and support scientific argumentation. The students, in a small collaborative working group of three, in these videos were asked to read several different alternative explanations for why ice melts at different rates when placed different types of materials, to determine which provided explanation was the most valid or acceptable and then to craft an argument in support of their chosen explanation. These videos of high schools students engaged in an episode of argumentation were collected as part of another study (Sampson & Clark, 2009, 2011).

The authors along with another rater viewed the videos together, but did not discuss their scores on each item until the video ended and each rater completed the protocol on his or her own. Scores for each item were then compared and when significant differences among the raters emerged, the item and the description of the item were discussed, evaluated, and modified in order to reduce ambiguity. This process resulted in numerous refinements to the stems and descriptions. An additional objective of this process was to identify and remove any items from the protocol that targeted an aspect of argumentation that was too difficult to observe or assess. This iterative refinement process was repeated over several cycles until the four raters were relatively consistent in their scoring. At the end of this step of the project, the protocol was reduced to a total of 20 items.

Step 6: A second expert review of the items. The observation protocol was sent out to the same expert reviewer group for further comments and to evaluate the translational and content validity of the revised items. Reviewer comments were critically considered in making adjustments to the text of the items, with particular attention paid to the descriptions for each stem as well as the inclusion of items in one of the four broad categories used to structure the protocol. This round of reviewer input resulted in responses from seven members of the panel (although some reviewers did not rate or comment on each item). Guided by these responses, the authors made several additional adjustments.

One item was deleted from the existing protocol due to its repetitive nature, as identified by reviewers, and agreed upon by the authors. Another item was also considered by the reviewers to be too similar to another specific stem, so the authors condensed those two items into one. Another structural change suggested by the reviewers, and agreed upon by the authors, was to combine two of the categorical aspects (conceptual and cognitive) and their items into one, more cohesive grouping. The resulting instrument contains 19 items that are divided into three categories (conceptual and cognitive aspects, epistemological aspects, and social aspects of scientific argumentation). The rating of the translational and content validity items by the panel of experts, along with the literature used to develop them, can be found in the *Results*.

Step 6: Analysis of the inter-rater reliability of the instrument. At this point, our focus moved from instrument development to the process of initial validation. We used the final version of the observation protocol from step 5 to score 20 different videos of students engaged in an episode of scientific argumentation during an actual lesson. All of the episodes took place during the "argumentation session" stage of either the *Argument-Driven Inquiry* (ADI) instructional model (Sampson & Gleim, 2009; Sampson, Grooms & Walker, 2009) or the *Generate an Argument* (GaA) instructional model (Sampson & Grooms, 2010). The argumentation session that is included in both of these instructional models is designed to give small groups of students an opportunity to propose, support, critique, and revise an evidence-based argument either by using data they collected through a method of their own design (ADI) or from a corpus of data provided to them (GaA).

Two of the authors served as raters for all 20 episodes of argumentation. The raters viewed the videos at the same time and recorded observation notes in the table provided as part of the ASAC protocol (see Appendix). Then, at the completion of each video, each rater assigned a score for each item on the protocol and recorded some of their observations to justify their decision. Once the raters had completed the protocol individually, the two raters compared their scores for each item. The score assigned by each rater on each item, as well as the total score assigned to each episode by the two raters, was recorded and then compared in order to evaluate the inter-rater reliability of the ASAC observation protocol (see Results).

Step 7: Analysis of the criterion-related validity of the instrument. To assess the convergent validity of the ASAC, which, as noted earlier, is an instrument's ability to provide a similar score to other instruments that are used to measure the same construct, we used the Toulmin Argument Pattern (TAP) framework (Erduran, Osborne & Simon, 2004; Osborne et al., 2004; Simon, Erduran & Osborne, 2006) to score a subset of 12 videos from the inter-rater reliably analysis. We then compared the ASAC and TAP scores for each episode in order to determine if the two methods resulted in similar conclusions about the overall quality of the scientific argumentation (see Results). We decided to use this framework to help validate the ASAC because it places a much greater emphasis on the structural components of an argument than the ASAC does; thus we predicted that there would be a strong correlation between ASAC scores and TAP rankings but that there would also be more variation in ASAC scores due to its more holistic focus.

In order to measure the quality of an episode of argumentation using the TAP framework (see Erduran, Osborne et al., 2004), researchers must first transcribe a discussion. In this case, we used the argumentation sessions we recorded during the various classroom lessons. The argumentative operations of each conversational turn are then coded using five different categories: (a) opposing a claim, (b) advancing claims, (c) elaborating on a claim, (d) reinforcing a claim with additional data and/or warrants, and (e) adding qualifications. One of these codes is applied to each conversational turn during the discussion that takes place during an episode of argumentation. Researchers must then identify the structural components of an argument (i.e., claim, counter-claim, data, warrants, rebuttals, etc.) that are found within and across the conversational turns.

After identifying the argumentative operations of each conversational turn and the various structural components of arguments voiced by the participants in the discussion, the quality of an argumentation episode is assessed using the hierarchy outlined in Table 12.1. The hierarchy is based on two major assumptions about what makes one episode of argumentation better than another. First, higher quality

Table 12.1 TAP argumentation quality hierarchy developed by Erduran, Simon, and Osborne (2004)

Quality	Characteristics of argumentation
Level 5	Extended arguments with more than one rebuttal.
Level 4	Arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counter-claims as well, but this is not necessary.
Level 3	Arguments with a series of claims or counter-claims with data, warrants, or backings with the occasional weak rebuttal.
Level 2	Arguments consisting of claims with data, warrants, or backings, but no rebuttals. Osborne advocates further distinction at this level: • Level 2B (2.5)—Arguments consisting of a claim supported by multiple pieces of data, warrants, or backings, but no rebuttals. • Level 2A (2.0)—Arguments consisting of a claim supported by a single piece of data, warrant, or backing, but no rebuttals.
Level 1	Arguments that are a simple claim versus a counter-claim or a claim versus claim.

argumentation must include arguments that consist of grounds (i.e., data, warrants, or backing) rather than unsubstantiated claims. Second, episodes of argumentation that include rebuttals (i.e., a challenge to the grounds used to support a claim) are "of better quality than those without, because oppositional episodes without rebuttals have the potential to continue forever with no change of mind or evaluation of the quality of the substance of an argument" (Erduran, Osborne, et al., 2004, p. 927).

Next, we used 20 videos from the inter-rater reliability analysis to evaluate the concurrent validity of the ASAC. Concurrent validity, as discussed earlier, concerns that ability of an instrument to discern between theoretically different groups. Therefore, if the ASAC has strong concurrent validity, we would expect an expert group of individuals (e.g., graduate students who understand the epistemological commitments, cognitive processes, and social norms that govern how people participate in scientific argumentation) to score higher on the ASAC than a group of novices (e.g., high school students who have little or no experience participating in scientific argumentation). We would also expect students to score higher on the ASAC after having numerous opportunities to participate in scientific argumentation.

The 20 videos were therefore divided into four theoretically distinct groups. The first group consists of high school students without any genuine experience with scientific argumentation ($n = 3$). The second group consists of undergraduate students who were video recorded as they participated in lab activity designed using the ADI instructional model for the first or second time at the beginning of a general chemistry lab course ($n = 7$). The third group also consists of undergraduate students; however, these videos were recorded at the end of a semester of general chemistry lab after the students had a chance to participate in four different ADI labs ($n = 7$). The final group consists of science education graduate students ($n = 3$). We then compared the average ASAC score of these four groups in order to determine if the ASAC could be used to distinguish between them as we predicted (see Results).

Results

In this section, we describe the results of our analysis of the reliability and validity of the ASAC in light of the methodological framework outlined in Fig. 12.1.

Construct validity. We evaluated the construct validity of the ASAC in two ways. During the construction of the first iteration of the ASAC protocol, the authors drew upon the argumentation literature to identify many common elements of what researchers and science educators considered to be characteristics of quality argumentation. To further strengthen the theoretical foundation of these items and their content validity, the authors relied on three important aspects of argumentation which, according to Duschl (2008), are fundamental to the process and make scientific argumentation different from the argumentation that takes place between individuals in other contexts. These aspects, as noted earlier, include: (1) the conceptual structures and cognitive processes used when reasoning scientifically; (2) the epistemic frameworks used when developing and evaluating scientific knowledge; and, (3) the social processes and contexts that shape how knowledge is communicated, represented, argued, and debated (p. 277).

We also evaluated the translational validity of the items and content validity of the instrument through two rounds of expert review. The comments and suggestions generated during this process assisted in shaping the wording and structure of the items, which, in turn, enhanced the translational validity of the instrument. The first iteration of the ASAC protocol underwent several major structural changes, resulting in an instrument comprised of 19 items measuring various aspects of quality argumentation, organized into three broad categories related to the theoretical framework. The reviewers' rating of the content and translational validity of each item from the second round of review as well as the empirical or theoretical foundation of each item is provided in Table 12.2. As illustrated in this table, the pool of expert reviewers agreed that each item is an important aspect of scientific argumentation and should be included in the instrument (min = 4.14/5, max = 5/5). The expert reviewers also agreed that the items, with the exception of items 8 and 15 (which were moved to a different section based on their feedback), were placed in the appropriate category and the corresponding section of the protocol (min = 4.17/5, max = 5/5).

The *Conceptual and Cognitive Aspects of Scientific Argumentation* section of the ASAC consists of seven items. These items allow a researcher to evaluate important elements of scientific argumentation, such as how much the participants focus on problem solving or advancing knowledge, how often individuals evaluate alternative claims, the participants' willingness to attend to anomalous data, the participants' level of skepticism, and the participants' use of appropriate or inappropriate reasoning strategies. The *Epistemic Aspects of Scientific Argumentation* section contains seven items. These items focus on important elements of scientific argumentation such as the participants' use of evidence, their evaluation of the evidence, the extent to which the participants use scientific theories, laws or models during the discussion, and how often the participants use the language of science to communicate their ideas. Finally, the *Social Aspects of Scientific Argumentation* section contains five items, which provides a means for assessing how the participants communicate and interact with each other. These items assess important elements of argumentation such as the participant's ability to be reflective about what they say, their respect for each other, their willingness to discuss ideas introduced by others, and their willingness to solicit ideas from others.

Inter-rater reliability. We used the scores from 20 different episodes of argumentation that were generated by two different raters to evaluate the inter-rater reliability of the ASAC. We first calculated a correlation coefficient between the two sets of total scores. The results of the analysis indicate that there was a significant and strong correlation between the scores of the two raters, $r(20) = 0.99$, $p < 0.001$. Figure 12.2 provides a scatter plot of the data points along with the equation for the best fitting line and the proportion of variance accounted for by that line. This estimate of the total score inter-rater reliability, $R^2 = 0.97$, is high for an observational protocol.

Table 12.2 The empirical and theoretical foundation of the items included in the ASAC along with the experts ratings of the content and translational validity of the items

		Important element[a]		Inclusion in the category[a]	
Items	Empirical or theoretical foundation	*M*	*SD*	*M*	*SD*
Conceptual and cognitive aspects					
• The work and talk of the group focused on the generation or validation of knowledge.[b]	Abell et al. (2000), Berland and Reiser (2009), Sandoval and Reiser (2004)	5.00	0.00	4.50	0.84
• The participants sought out and discussed alternative claims or explanations.	Driver et al. (2000), Duschl and Osborne (2002), Sampson and Clark (2011)	4.83	0.41	5.00	0.00
• The participants modified their claim or explanation when they noticed an inconsistency or discovered anomalous information.	Berland and Reiser (2009), Duschl and Osborne (2002), Vellom and Anderson (1999)	5.00	0.00	4.80	0.45
• The participants were skeptical of ideas and information.	Driver et al. (2000), Osborne et al. (2004)	4.71	0.49	4.57	0.79
• The participants provided reasons when supporting or challenging an idea.	Erduran and Jimenez-Aleixandre (2007)	5.00	0.00	4.71	0.76
• The participants based their decisions or ideas on inappropriate reasoning strategies.[c]	Duschl (2008), Vellom and Anderson (1999), Zeidler (1997)	4.57	0.79	4.33	0.98
• The participants attempted to evaluate the merits of each alternative explanation or claim in a systematic manner.	Duschl (2008), Osborne et al. (2004)	4.86	0.38	4.43	0.79
Epistemological aspects					
• The participants relied on the "tools of rhetoric" to support or challenge ideas.[c,d]	Kuhn (1993), Zeidler (1997)	4.14	0.90	3.29	1.50
• The participants used evidence to support and challenge ideas or to make sense of the phenomenon under investigation.	Berland & Reiser (2009), Duschl (2008), Erduran and Jimenez-Aleixandre (2007)	5.00	0.00	4.43	0.79
• The participants examined the relevance, coherence, and sufficiency of the evidence.	Driver et al. (2000), Duschl (2007)	5.00	0.00	4.17	1.33
• The participants evaluated how the available data was interpreted or the method used to gather the data.	Duschl (2007), Duschl and Osborne (2002)	4.71	0.76	4.57	0.79

Table 12.2 (continued)

Items	Empirical or theoretical foundation	Important element[a]		Inclusion in the category[a]	
		M	*SD*	*M*	*SD*
• The participants used scientific theories, laws, or models to support and challenge ideas or to help make sense of the phenomenon under investigation.	Driver et al. (2000), Duschl (2008), Sandoval (2003)	5.00	0.00	4.43	0.98
• The participants made distinctions and connections between inferences and observations explicit to others.	Driver et al. (2000), Erduran and Jimenez-Aleixandre (2007), Sandoval (2003)	4.57	0.54	4.43	0.98
• The participants used the language of science to communicate ideas.	Carlsen (2007), Erduran and Jimenez-Aleixandre (2007)	4.57	0.54	4.00	1.16
Social aspects					
• The participants were reflective about what they know and how they know.[e]	Alexopoulou and Driver (1996), Erduran and Jimenez-Aleixandre (2007)	4.50	0.55	3.83	0.98
• The participants respected what each other had to say.	Boulter and Gilbert (1995), Richmond and Striley (1996)	4.43	1.13	4.71	0.76
• The participants discussed an idea when it was introduced into the conversation.	Berland and Reiser (2009), Sampson & Clark (2011)	4.86	0.38	4.86	0.38
• The participants encouraged or invited others to share or critique ideas.	Boulter and Gilbert (1995), Sampson and Clark (2011)	4.71	0.76	5.00	0.00
• The participants restated or summarized comments and asked each other to clarify or elaborate on their comments.	Alexopoulou and Driver (1996), Boulter and Gilbert (1995), Richmond and Striley (1996)	5.00	0.00	5.00	0.00

[a] The items were scored using the following scale: 1—strongly disagree, 2—disagree, 3—neutral, 4—agree, 5—strongly Agree.

[b] Item was reworded to read, "The conversation focused on the generation or validation of claims or explanations" based on reviewer feedback.

[c] Item represents an undesirable element of scientific argumentation and is therefore scored in reverse.

[d] Item was moved from the conceptual and cognitive aspects section of the protocol to the epistemological aspects section based on expert feedback.

[e] Item was moved from the conceptual and cognitive aspects section of the protocol to the social aspects section based on expert feedback.

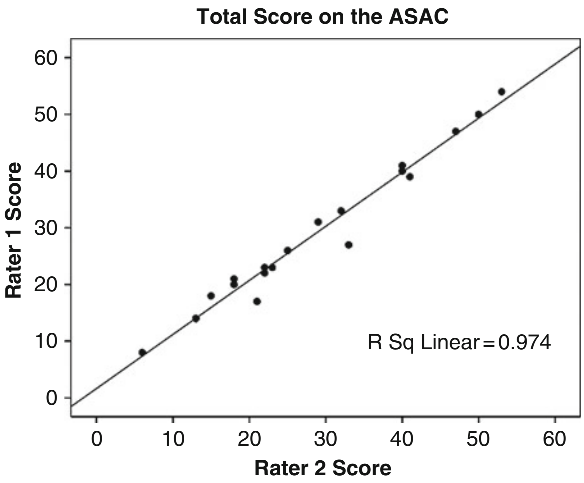

Fig. 12.2 A scatter plot of the total ASAC score generated by two different raters

Next, we calculated a correlation coefficient between the two sets of scores for the items. The results of this analysis, once again, indicate that there was a significant and strong correlation between the scores of the two raters, $r(380) = 0.89$, $p < 0.001$. Figure 12.3 provides a scatter plot of the 380 data points along with the equation for the best fitting line and the proportion of variance accounted for by that line ($R^2 = 0.795$). This estimate of the inter-rater reliability for the items is also high (although not as high as for the total score). The scatter plot in Fig. 12.3 also illustrates that most of the observed discrepancies between the scores produced by the two raters for the various items were with a single point. This indicates that even

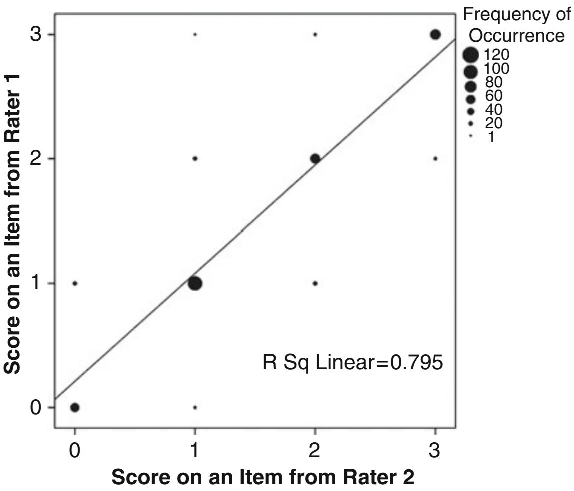

Fig. 12.3 A scatter plot of the scores on each item generated by two different raters

when the raters did not assign the same score on an item their evaluation was at least similar. It is important to note, however, that this type of analysis does not take into account chance agreement between the two raters. The influence of chance agreement on estimates of inter-rater reliability is important to consider when a scale is used because a scale limits choice and can artificially inflate the measure.

We therefore also used this corpus of data to calculate a Cohen's Kappa value as an additional estimate of inter-rater reliability. Cohen's Kappa, which is often used to measure the consistency and repeatability of scores between two raters, takes into account chance agreement unlike correlations or measures of percent agreement. The maximum value of Kappa is 1.0, which indicates perfect agreement, and a value of 0.0 indicates that the observed agreement is the same as that expected by chance. Landis and Koch (1977) suggest that values of Kappa above 0.60 indicate "good to excellent" agreement between scores of two raters, and values of 0.40 or less show "fair to poor" agreement. The inter-rater reliability of the ASAC, as measured with Cohen's Kappa, was 0.69. Thus, this analysis indicates that two different raters can use the ASAC to score an episode of argumentation in a manner that is consistent and repeatable.

Criterion-related validity. As a first test of this important aspect of validity, we compared ASAC scores from 12 episodes of argumentation to the scores we generated using the TAP framework. We predicted, as discussed earlier, that these two approaches would result in similar conclusions because they both are designed to measure the same construct. We therefore calculated a correlation coefficient between the two sets of total scores. The results of the analysis indicate that there was a significant and strong correlation between the two sets of scores, $r(12) = 0.96$, $p < 0.001$. Figure 12.4 provides a scatter plot of the 12 episodes of argumentation that we scored using both measures. As illustrated in this figure, the videos of the high school students engaged in argumentation received a total ASAC score in the range of 8–22 (out of a possible 57) and a TAP ranking of Level 2B (2.5 out of a possible 5). The videos of the undergraduate students were given a total ASAC score in the range of 7–33 and TAP rankings between 2.5 and 4. The videos of the graduate students, in contrast with the other two groups, received total ASAC scores in the range of 47–53.5 and all earned a Level 5 TAP ranking. Overall, these findings suggest that the ASAC and the TAP framework measure the same underlying construct and the ASAC, as a result, has adequate convergent validity.

We then examined how well the instrument is able to distinguish between groups that it should be able to distinguish between based on our theoretical framework (i.e., students with different levels of knowledge and skills needed to participate in scientific argumentation). A Kruskal–Wallis test (a nonparametric alternative to an ANOVA) was conducted to evaluate differences among the four groups (high school students, undergraduate students at the beginning of the semester, undergraduate students at the end of the semester, and graduate students) on the median ASAC score. The test, which was corrected for tied ranks, was significant, $\chi^2\,(3, N = 20) = 10.88$, $p = 0.01$. Follow-up tests were conducted to evaluate pairwise difference among the four theoretical groups, controlling for type I error across the tests by using the Holm's sequential Bonferroni approach. The results of these tests indicate

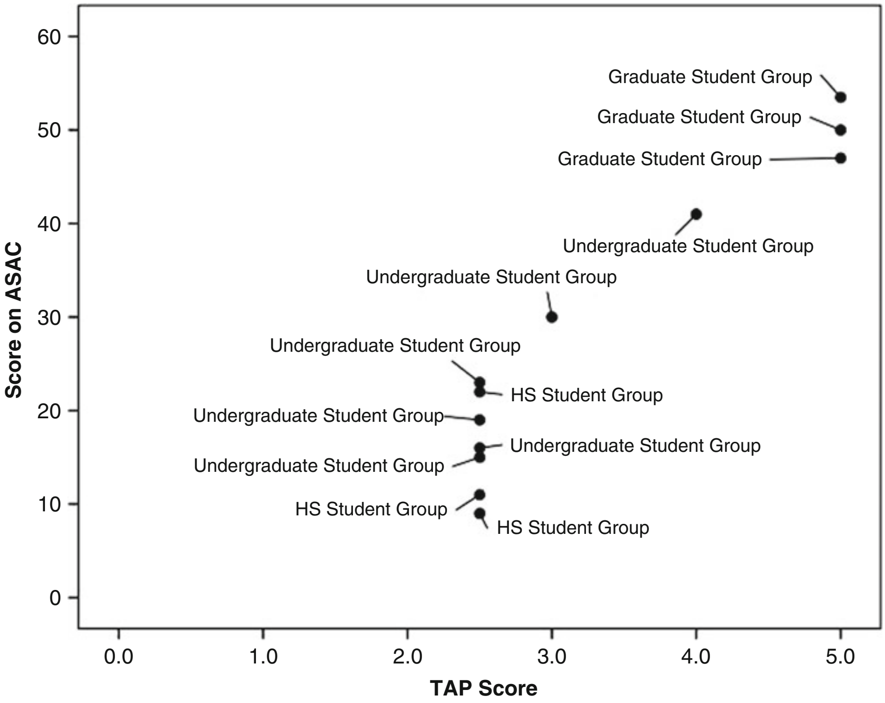

Fig. 12.4 Comparison of ASAC scores and TAP rankings

a significant difference in the median ASAC scores between the high school students and the graduate students, the undergraduate student group at the beginning of the semester and the graduate students, and the undergraduate students at the end of the semester and the graduate students. Figure 12.5 provides a boxplot of the distribution in ASAC scores for the four different groups. Overall, this analysis indicates that the instrument can distinguish between groups of students who should have different argumentation skills as expected and therefore has adequate concurrent validity.

Discussion and Limitations

The development of an ASAC observation protocol serves as another contribution to the expanding field of research that examines argumentation in science education. As the importance of developing students' ability to engage in productive scientific argumentation continues to grow, researchers and educators must have methods for assessing how students participate in this complex practice. Several frameworks have been developed in recent years that focus on the structure of the arguments produced by students during an episode of argumentation in order to help fulfill

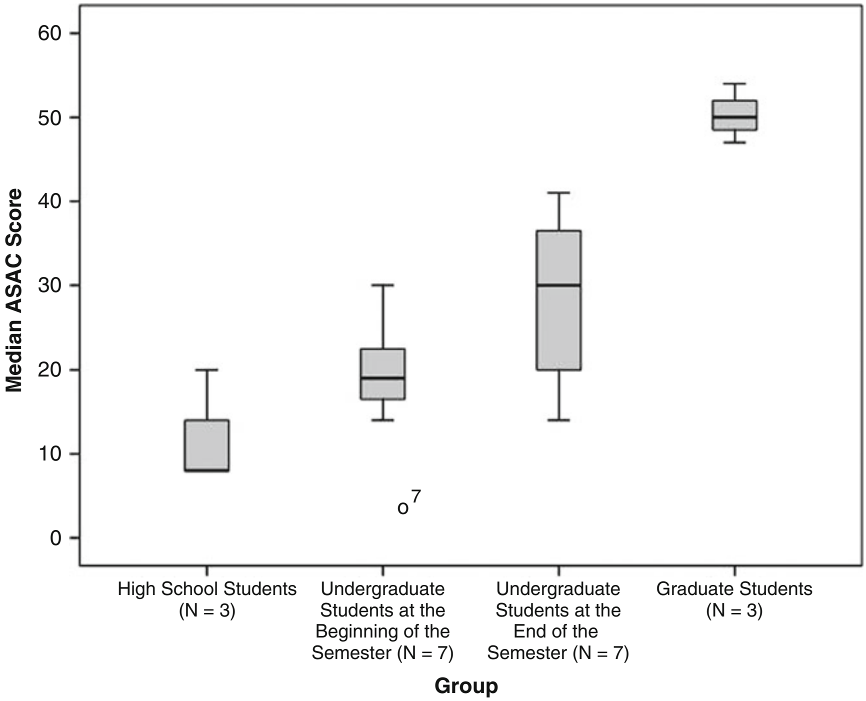

Fig. 12.5 Distribution of ASAC scores for different groups of students

this need (Berland & Reiser, 2009; Erduran, Osborne et al., 2004; McNeill, Lizotte, Krajcik & Marx, 2006; Sampson & Clark, 2008; Sandoval, 2003). These approaches have aptly focused on the general structure of arguments and their justifications, the content validity of these products, and the epistemological appropriateness of them. However, there has been less work that emphasizes and investigates another critical component, the social process and activities involved in generating those arguments.

The ASAC observation protocol, as developed and initially validated through this study, should serve as a useful tool to incorporate all of the aforementioned elements as well as bringing the social activities involved in the process into consideration. The research presented here helps to establish the construct validity of the instrument by sharing the literature that was used to define the construct of scientific argumentation and develop the items used to measure it. The construct validity was further supported through the guidance and critique provided during instrument construction by a panel of several expert reviewers comprised from science education researchers with a research record in this topic. These stages of validation support the assertion that the ASAC observation protocol does measure the construct of *scientific argumentation*.

Likewise, the initial criterion-related validity of the ASAC, which is the ability of the instrument to provide scores expected on theoretical grounds, was also established through the several different approaches. The ASAC, when used to assess a collection of videos of students engaged in structured activities, resulted in scores that varied substantially between groups of students that theoretically should differ in argumentation skills. The authors also used the TAP framework to help demonstrate that the ASAC and TAP both measure argumentation quality. Finally, inter-rater reliability, an assessment of the possibility for researchers to produce similar and repeatable scores, was demonstrated through several quantitative procedures. The satisfactory Cohen's Kappa value obtained along with the strong and significant correlations between two raters' sets of scores provide evidence that two raters can produce a similar, if not identical, assessment of an episode of argumentation when using the ASAC.

The authors, however, readily acknowledge that the ASAC protocol presented here and the validation measures described represent a first attempt to develop a research instrument that can measure conceptual, cognitive, epistemological, and social aspects of argumentation that emerge during classroom activities. Although the data set used during this research provides sufficient evidence for the validity and reliability of the ASAC protocol, additional data should be collected to further strengthen the quantitative data set and related measures. Increasing the amount of scores from further video reviews offers the potential to demonstrate further inter-rater reliability and an increase in Kappa and correlation values. Another benefit from an increased data set would be the capability to analyze each item in the instrument using Kappa calculations. At this stage of our research, the data set was not large enough to allow for these more detailed assessments. Increased data collection using the ASAC protocol could also increase the variety of contexts to which the instrument is applied, beyond the structured "argumentation session" that were used as a source of data in this study.

It is also important to note that the videos of the activities used in the data set reported here represent a particular type of approach for promoting and supporting student engagement in scientific argumentation, one that is structured to encourage students to share, critique, and refine evidence-based arguments that provide an answer to a research question. The authors concede that using this type of activity does provide a minimal amount of forced argumentation (because the argumentation sessions are design to foster it), excluding the validity of the instrument for more naturally emerging instantiations of classroom-based argumentation. However, we feel that this limitation should be considered to be an invitation to extend the use of the ASAC protocol to a variety of contexts, particularly in light of its unique and contextually bound social component.

The context of an episode of argumentation, especially the structure of the activity used to engage students in this practice, will influence the magnitude of the score on the ASAC (or the score on any instrument for that matter). For example, any classroom activity that does not allow students to collaborate with each other would not be able to score as well on the instrument, and a classroom culture that discourages verbal interactions between students will also result in lower scores. Another contextual aspect of the activity that might influence ASAC scores is an opportunity for students to generate their own data and explanations. We noticed during our review of the videos that the groups that were provided with a collection of data and a list of explanation to select from were not as likely to engage in extended episode of argumentation. Instead, these students seemed to select an explanation and then search for confirming evidence. In activities where students generated their own data, students had to make sense of it and develop their explanations, offering more opportunities to engage in scientific argumentation. Thus, the nature of the activity must be deeply considered when raters use the ASAC protocol to assess argumentation quality or to make comparisons across groups. However, we feel that this issue is a potential strength of the instrument; it will allow researchers to examine how the structure of activity influences the way students engage in scientific argumentation.

Another implication for using the ASAC protocol that can be noted from this study involves the necessity of familiarization and training the raters that will use the instrument. The authors of this chapter, who also served as raters, were involved with every step of the instrument development; therefore they were rather familiar with the content of the protocol. However, even in light of knowledge of the instrument, the two raters still had to refresh their understandings and align their interpretation of the items before beginning a scoring session. We found that watching a few practice videos, scoring them, and then discussing the discrepancies in scores were all that was needed to "calibrate" the raters. Raters should complete at least two "trail runs" by watching and scoring videos of an activity similar to the one that will be assessed before scoring the actual data set in order to help ensure the highest possible inter-rater reliability.

In conclusion, the ASAC observation protocol should serve educational researchers well, as investigations into the benefits and effectiveness of argumentation in science classrooms continues to grow. This growth has potential to move into other areas of concern, such as gender and cultural dynamics that can influence the process and the product. The ability for researchers to measure quality over periods of time within specific groups or environments can be facilitated through this instrument. This use can allow researchers and teachers alike to measure the progress of change in students' abilities to engage in productive scientific argumentation and enhancement of scientific argumentation skills. Thus, the ASAC instrument, although nascent in its development, offers a much-needed tool to help researchers understand and science education realize some of the visions of reform and literacy underpinning many current efforts.

Recommendations for Using the ASAC

The ASAC, as noted earlier, is criterion-referenced assessment of the quality of an episode of argumentation. The observers' judgments should therefore not reflect a comparison with any other instructional setting or event. The instrument contains 19 items. Seventeen items are rated on a scale from 0 (not at all) to 3 (often) and two items, which target undesirable aspects of scientific argumentation, are rated in reverse (0—often, 3—not at all). Possible scores range from 0 to 57 points, with higher scores reflecting higher quality scientific argumentation. The ASAC can be used in a wide range of educational contexts and levels including middle schools, high schools, and universities. It can be used to score a live event (e.g., to score an episode of argumentation as it unfolds in the classroom, to score the nature of argumentation inside a classroom where permission to video record students has not been granted, etc.) or a video recording of a past event.

An observer should adhere to the following procedure when using the ASAC (assuming that the observer has been trained about how to use the instrument and has completed several trail runs with another rater to ensure that his or her interpretations of the items are aligned with the content of the item descriptions). First, the observer should turn to the *record of events* section (see Appendix) and take observational notes in the provided table while watching the entire episode of argumentation. After the episode is complete, the observer should then turn the section of the protocol with the 19 items and score them. The observer should also include observations they made in the space for *comments* under each item description in order to support his or her ranking of an item. Finally, the rater should return to the sections called *observational information*, *group characteristics*, and *activity design* and fill in all the necessary background information based on his or her observations of the episode of argumentation and, if necessary, ask the classroom teacher to provide any additional information that is needed.

We, however, recommend that only trained observers use the ASAC. Although the protocol includes a detailed description of the aim of each item, raters need to participate in a formal training program. This training program, at a minimum, should include an opportunity for the trainees to examine the content and aim of each item, observe videotapes of episodes of argumentation or an actual instance of argumentation occurring in a classroom, score them using the ASAC, and discuss their interpretations of the items with others. As part of this process, raters should be encouraged to review videos together and discuss discrepancies in order to bring their personal interpretations of the items into alignment with the actual content of the item descriptions (e.g., raters tend to disregard an aspect of a description or interpret the content of an item description in unintended ways). We also recommend, as noted earlier, that raters watch several trail videos, score them, and discuss any discrepancies in order to "calibrate" with the instrument and each other before beginning a scoring session associated with a research study.

The ASAC can be used for several different research purposes. First, it can be used in longitudinal studies to examine how students' ability to participate in scientific argumentation changes over time. Researchers, for example, can use the ASAC to assess the quality of a series of subsequent events that provide students with an opportunity to engage in an episode of argumentation in order to determine how much students improve as a result of a new curriculum or new instructional method over the course of a semester or a school year. Second, the instrument can be used in comparison studies to examine the efficacy of a new curriculum or instructional strategy as a way to improve scientific argumentation skills. It can also be used to compare various designs of a new instructional method or ways of organizing the structure of an activity as part of development project. Researchers, for example, might be interested in comparing the nature of the quality of the argumentation that takes place between students when they are required to generate and make sense of their own data as part of the activity versus being supplied with an existing data (which we discussed earlier). Researcher can also use the ASAC to determine gains in argumentation skills in an experimental or quasi-experimental research study that uses a performance task as a pre–post intervention assessment.

Science teacher educators can also use the ASAC for professional development purposes because science teachers often do not know “what to look for and how to guide their students’ arguments” (Driver et al., 2000, p. 295) or how to monitor their students’ progress as they learn how to participate in scientific argumentation. We think the ASAC will help teachers with this difficult task. Science teacher educators, for example, can train science teachers to use the observational protocol which, in turn, would help science teachers develop a better understanding of what counts as high-quality scientific argumentation (i.e., increase their understanding of scientific argumentation). Once trained, these teachers could then use the ASAC in their own classrooms to assess how well their students participate in argumentation. These teachers could then use the information they gathered using the ASAC to guide their own classroom practice and to plan future lessons. Thus, the ASAC should provide science teachers, science teacher educators, and science education researchers with a valid and reliable way to assess the quality of argumentation, so better curricular and instructional decision can be made about what works and what needs to be fixed.

Acknowledgements The development and validation of the instrument reported here was supported, in part, by the Institute of Education Science, U.S. Department of Education through grant #R305A100909 to Florida State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Appendix

ASSESSMENT OF SCIENTIFIC ARGUMENTATION IN THE CLASSROOM OBSERVATION PROTOCOL

OBSERVATION INFORMATION

Teacher: ______________________ School: ______________________

Subject: ______________________ Grade: ______________________

Observer: ______________________ Date: ______________________

Duration of the episode: ______________________________________

GROUP CHARACTERISTICS

Size:
□ 2
□ 3
□ 4
□ 5
□ 6 or More
□ Whole Class

Number of times that these students have been placed into this same group before:
□ Never
□ 1
□ 2
□ 3
□ 4 or more
□ Unknown

Assignment to the Group:
□ Random
□ Self-Selected
□ Achievement – Mixed
□ Achievement – High
□ Achievement – Low
□ Teacher choice – Other
□ Unknown

Gender Composition:
□ All Male
□ All Female
□ # of Males > # of Females
□ # of Females > # of Males
□ # of Females = # Males

Racial/Ethnic Composition: ______________________

Native Language Composition: ______________________

ACTIVITY DESIGN

Provide a brief description of (a) the way the activity or lesson was designed in an effort to promote and support argumentation and (b) the way the teacher encouraged students to engage in argumentation.

RECORD OF EVENTS

In the space provided keep a running record of the events that occurred as the participants interacted with each other, the materials, and ideas.

Time	Description of Event

CONCEPTUAL AND COGNITIVE ASPECTS OF SCIENTIFIC ARGUMENTATION

How the group attempts to negotiate meaning or develop a better understanding

1. The conversation focused on the generation or validation of claims or explanations.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: The emphasis on the generation or validation of claims or explanations indicates that there were some significant claims or explanations at the heart of discussion. Groups that score high on this item maintain the focus of their talk and efforts on understanding or solving the problem rather than the best way to finish their work quickly or with the least amount of effort. *Note:* Groups that stay on topic but never go engage in an in-depth discussion about what is happening should be scored low on this item.				
Comments:				

2. The participants sought out and discussed alternative claims or explanations.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: Divergent thinking is an important part of scientific argumentation. A group that meets this criterion would talk about more than one claim, explanation, or solution. Individuals that valued alternative modes of thinking would respect and actively solicit new or alternative claims, explanations, or solutions from the other participants. *Note:* Groups that discuss multiple types of grounds or support for a claim, explanation, or solution but only one claim, explanation, or solution should be scored low on this item.				
Comments:				

3. The participants modified their claim or explanation when they noticed an inconsistency or discovered anomalous information.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: Inconsistencies between claims or explanation and the phenomenon under investigation are common in science. A group that modified their claim or explanation when they noticed inconsistencies or anomalies would not ignore “things that do not fit” or attempt to discount them once they are noticed by one of the participants. Groups that score high on this item try to modify their claim or explanation (not just their reasons) in order to account for an inconsistency or an anomaly rather than attempting to “explain them away” or simply deciding that something “doesn’t matter.”				
Comments:				

4. The participants were skeptical of ideas and information.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: During scientific argumentation, allowing a variety of ideas to be presented, but insisting that challenge and negotiation also occur would indicate that group members were skeptical. Accepting ideas without accompanying reasons would result in a low score because it is a sign of credulous thinking. In other words, students must be willing to ask, “how do you know?” or “Are you sure?” Groups that respond to the ideas of others with comments such as “ok”, “that sounds good to me”, or “whatever you think is right” would score low on this item.				

Comments:

5. The participants provided reasons when supporting or challenging an idea.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often

Description: Providing reasons to support or challenge a claim, conclusion, or explanation is a crucial characteristic of argumentation. Claims must have some support provided for them beyond simply restating the claim itself. Making claims with out support would result in a low score on this item and including any reason like "that's what I think", "it doesn't make sense", "the data suggests…" or "but that doesn't fit with…" would result in a higher score. *Note:* Personal or past experiences count as a reason for this item.

Comments:

6. The participants based their decisions or ideas on inappropriate reasoning strategies.	**3**	**2**	**1**	**0**
	Not at all	Once or Twice	A few times	Often

Description: When people are trying to support ideas they often: (a) jump to hasty generalizations, (b) attribute causality to random events, (c) insist that a correlation is evidence of causality, and (d) exhibit a confirmation bias (for example saying, "now we need some data to prove this"). Groups that avoid inappropriate reasoning strategies or recognize them when they occur would score high on this item. Groups where these types of reasoning strategies are common would score low on this item.

Comments:

7. The participants attempted to evaluate the merits of each alternative explanation or claim in a systematic manner.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often

Description: This addresses the tentative or responsive nature of science. The idea that there is often more than one way to interpret data or evidence and that only through careful analysis can an idea be accepted or eliminated. This gets at the "gut" response factor. Conclusions are not based on opinion or inference.

Comments:

EPISTEMIC ASPECTS OF SCIENTIFIC ARGUMENTATION

How consistent the process is with the culture and norms of science

8. The participants relied on the "tools of rhetoric" to support or challenge ideas.	**3**	**2**	**1**	**0**
	Not at all	Once or Twice	A few times	Often

Description: "Tools of rhetoric" refer to tricks or strategies used to win a debate. Tool of rhetoric include: (a) claiming that if someone cannot disprove a claim it must be true, (b) using emotive words and false analogies, (c) directing the focus of the discussion from thinking about a claim or an explanation to thinking about the person holding or proposing a claim or an explanation, (d) over-relying on authorities, (e) dichotomizing issues so that if you discredit one position, then the observer is forced to accept the other view, and (f) making claims that are a simple restatement of one of the premises. Groups that avoided using the tools of rhetoric would score high on this item. *Note:* This item focuses on how the content of a discussion is presented or supported (i.e., how they are saying it) rather than the content of the discussion (i.e., what they are saying).

Comments:

9. The participants used evidence to support and challenge ideas or to make sense of the phenomenon under investigation.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often

Description: A goal of scientific argumentation is the use of data as evidence to defend a claim, conclusion, or explanation. This item implies that students were attempting to use evidence in their arguments. This should more than an opinion; they must include data. Statements like "that's what I think" or "it doesn't make sense" would result in a low score. Statements like "the data we found suggests that …" or "our evidence indicates…" would result in a higher score.

Comments:

10. The participants examined the relevance, coherence, and sufficiency of the evidence.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often

Description: This item draws attention to the amount and kinds of evidence used to support a claim or explanation. Groups that attempt to (a) determine the value of a piece of evidence (e.g., "does that matter?"), (b) look at links or the relationship between multiple pieces of evidence (e.g., "This supports X and Y but this only supports X"), or (c) attempt to determine if there is enough evidence to support an idea (e.g., "We do not have any evidence to support that") would score higher on this item.

Comments:

11. The participants evaluated how the available data was interpreted or the method used to gather the data.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: The evidence provided for a claim or explanation should be evaluated based on how well the data was gathered and interpreted. A question such as "Why is that evidence included?" or "How did they gather their data?" or "Where did that data come from?" indicates that the participants are assessing methods or an interpretation of data and would result in a higher score.				
Comments:				

12. The participants used scientific theories, laws, or models to support and challenge ideas or to help make sense of the phenomenon under investigation.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: Science is theory-laden. In other words, scientists rely on broad, well-supported organizing ideas to frame their arguments and claims. Students should also employ these paradigmatic ideas in providing warrants for the evidence and claims they make or use to refute others' claims. Explicit reference to these "big ideas" will result in a higher score on this item.				
Comments:				

13. The participants made distinctions and connections between inferences and observations explicit to others.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: The structure of scientific arguments includes evidence involving both empirical (such as quantitative measurements and systematic observations) and inferential (noting of trends and logical connections among observations) aspects. Making these distinctions and their connections explicit to others enhances the quality of the argumentation and thus results in a higher score.				
Comments:				

14. The participants used the language of science to communicate ideas.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: This item stresses the importance of the accurate use of scientific language by students. The adoption and use appropriate terms (e.g., condensation, force, etc), phrases (e.g., "it supports" rather than "it proves") or ways of describing information is a characteristic of argumentation that is scientific. *Note:* Ideas may be explicated before being labeled with the correct terminology.				
Comments:				

SOCIAL ASPECTS OF SCIENTIFIC ARGUMENTATION

How the participants interact with each other

15. The participants were reflective about what they know and how they know.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: It is important for members of the group to agree on what they know and to be specific about how they know. Statements such as, “do we all agree?” or “is there anything else we need to figure out?” or “can we be sure?” indicate that participants are monitoring their progress and have an end goal in mind.				
Comments:				
16. The participants respected what each other had to say.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: Respecting what others have to say is more than listening politely or giving tacit agreement. Respect also indicates that what others had to say was actually heard and considered (e.g., “that is a good point”, interesting idea”, or “I hadn’t thought of that”). A group that scored high on this would allow everyone to present their ideas and express their opinions without censure or ridicule.				
Comments:				
17. The participants discussed an idea when it was introduced into the conversation.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: To be a participating and contributing member of the group, it is important to feel valued. Ideas and opinions need to be critically acknowledged. This means they are considered and given weight by the group. Groups that ignore ideas when they are proposed (results in the same idea being mentioned over and over) would earn a low score on this item.				
Comments:				
18. The participants encouraged or invited others to share or critique ideas.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: Good argumentation comes from considering and comparing competing ideas from multiple individuals to construct the most robust explanation of the phenomenon under study. Groups that consist of individuals that invite others to share (e.g., “what do you think”), critique (e.g., “do you agree” or “it is ok to disagree with me”), or discuss an idea (e.g., “let’s talk about this some more”) would score higher that a group with an alienating leader that dominates the conversation and the work of the group.				

Comments:				
19. The participants restated or summarized comments and asked each other to clarify or elaborate on their comments.	**0**	**1**	**2**	**3**
	Not at all	Once or Twice	A few times	Often
Description: The depth of discussion will be enhanced by not making implicit judgments or assumptions about another person's ideas or views, and it demonstrates that their point of view is valued and is furthering the discussion. Communication provides students with opportunities to identify the strengths and weaknesses of their understanding.				
Comments:				

Total: /57

References

Abell, S. K., Anderson, G., & Chezem, J. (2000). Science as argument and explanation: Exploring concepts of sound in third grade. In J. Minstrell & E. H. Van Zee (Eds.), *Inquiry into inquiry learning and teaching in science* (pp. 100–119). Washington, DC: American Association for the Advancement of Science.

Alexopoulou, E., & Driver, R. (1996). Small-group discussion in physics: Peer interaction modes in pairs and fours. *Journal of Research in Science Teaching, 33*(10), 1099–1114.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Andriessen, J., Baker, M., & Suthers, D. (2003). Argumentation, computer support, and the educational contexts of confronting cognitions. In J. Andriessen, M. Baker, & D. Suthers (Eds.), *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments* (pp. 1–25). Dordrecht: Kluwer.

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

Borg, W. R., & Gall, M. D. (1989). *Educational research: An introduction* (5th ed.). White Plains: Longman.

Boulter, C. J., & Gilbert, J. K. (1995). Argument and science education. In P. J. M. Costello & S. Mitchell (Eds.), *Competing and consensual voices: The theory and practices of argument* (pp. 84–98). Clevedon: Multilingual Matters.

Burns, R. B. (1994). *Introduction to research methods in education*. Chesire: Longman.

Carlsen, W. S. (2007). Language and science learning. In S. K. Abell & N. Lederman (Eds.), *Handbook of research on science education* (pp. 57–74). Mahwah, NJ: Lawrence Erlbaum Associates.

deVries, E., Lund, K., & Baker, M. (2002). Computer-mediated epistemic dialogue: Explanation and argumentation as vehicles for understanding scientific notions. *Journal of the Learning Sciences, 11*(1), 63–103.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287–313.

Duschl, R. (2007). Quality argumentation and epistemic criteria. In S. Erduran & M. Jimenez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research* (pp. 159–175). Dordrecht: Springer Academic.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education, 32*, 268–291.

Duschl, R., Schweingruber, H., & Shouse, A. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.

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

Edwards, A. (1957). *Techniques for attitude scale construction*. New York: Appleton-Century-Crofts.

Erduran, S. (2007). Methodological foundations in the study of argumentation in science classrooms. In S. Erduran & M. Jimenez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research.* (pp. 47–69). Dordrecht: Springer Academic.

Erduran, S., & Jimenez-Aleixandre, M. (Eds.). (2007). *Argumentation in science education: Perspectives from classroom-based research*. Dordrecht: Springer Academic.

Erduran, S., Osborne, J., & Simon, S. (2004). The role of argument in developing scientific literacy. In K. Boersma, O. deJong, H. Eijkelhof, & M. Goedhart (Eds.), *Research and the quality of science education* (pp. 381–394). Dordrecht: Kluwer.

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.

Kelly, G. J., Druker, S., & Chen, C. (1998). Students' reasoning about electricity: Combining performance assessments with argumentation analysis. *International Journal of Science Education, 20*(7), 849–871.

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

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

Kuhn, L., & Reiser, B. (2005). *Students constructing and defending evidence-based scientific explanations.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, TX.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

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.

National Research Council. (1999). *Transforming undergraduate education in science, mathematics, engineering, and technology*. Washington, DC: National Academies Press.

National Research Council. (2005). *America's lab report: Investigations in high school science*. Washington, DC: National Academies Press.

National Research Council. (2008). *Ready, set, science: Putting research to work in K-8 science classrooms*. Washington, DC: National Academies Press.

National Science Foundation. (1996). *Shaping the future: Strategies for revitalizing undergraduate education.* Paper presented at the National Working Conference, Washington, DC.

Nunally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in science classrooms. *Journal of Research in Science Teaching, 41*(10), 994–1020.

Richmond, G., & Striley, J. (1996). Making meaning in the classroom: Social processes in small-group discourse and scientific knowledge building. *Journal of Research in Science Teaching, 33*(8), 839–858.

Rubba, P. A., & Anderson, H. O. (1978). Development of an instrument to assess secondary school students: Understanding of the nature of scientific knowledge. *Science Education, 62*(4), 449–458.

Sampson, V., & Clark, D. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*(3), 447–472.

Sampson, V., & Clark, D. (2009). The effect of collaboration on the outcomes of argumentation. *Science Education, 93*(3), 448–484.

Sampson, V., & Clark, D. (2011). A comparison of the collaborative scientific argumentation practices of two high and two low performing groups. *Research in Science Education, 41*(1), 63–97.

Sampson, V., & Gleim, L. (2009). Argument-driven inquiry to promote the understanding of important concepts and practices in biology. *American Biology Teacher, 71*(8), 471–477.

Sampson, V., & Grooms, J. (2010). Generate an argument: An instructional model. *The Science Teacher, 77*(5), 33–37.

Sampson, V., Grooms, J., & Walker, J. (2009). Argument-driven inquiry: A way to promote learning during laboratory activities. *The Science Teacher, 76*(7), 42–47.

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

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

Simon, S., Erduran, S., & Osborne, J. (2006). Learning to teach argumentation: Research and development in the science classroom. *International Journal of Science Education, 28*(2&3), 235–260.

Suthers, D. (2006). Technology affordances for intersubjective meaning making: A research agenda for CSCL. *International Journal of Computer Supported Collaborative Learning, 1*(3), 315–337.

Trochim, W. M. (1999). *The research methods knowledge base* (2nd ed.). Cincinnati, OH: Atomic Dog.

Veerman, A. (2003). Constructive discussions through electronic dialogue. In J. Andriessen, M. Baker, & D. Suthers (Eds.), *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments* (pp. 117–143). The Netherlands: Kluwer.

Vellom, R. P., & Anderson, C. W. (1999). Reasoning about data in middle school science. *Journal of Research in Science teaching, 36*(2), 179–199.

Zeidler, D., L. (1997). The central role of fallacious thinking in science education. *Science Education, 81*, 483–496.