Scientific Argumentation in an Online AP Physics 1 Course

by

John Robert Kernion

B. S. in Ceramic Science and Engineering, The Pennsylvania State University, 1981

M. A. in Liberal Studies, Duquesne University, 1999

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Education

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

John Robert Kernion

It was defended on

June 17, 2019

and approved by

Kari Kokka, Assistant Professor, Instruction and Learning

Andy Cavagnetto, Associate Professor, Teaching and Learning, Washington State University

Dissertation Advisor: Ellen Ansell, Associate Professor, Instruction and Learning

Scientific Argumentation in an Online AP Physics 1 Course

John Robert Kernion, EdD

University of Pittsburgh, 2019

If one were to list a group of skills essential for scientific literacy on which the educational research community finds consensus, argumentation would undoubtedly be included, and perhaps be the top choice. This understanding has recently prompted the College Board to adapt its algebra-based AP Physics 1 course expectations by making the development of scientific argumentation skill an integral aspect of successful instruction. Although computational skill development and conceptual understanding of physics are still the backbone of the course, process skills such as argumentation have now taken on significant importance and must be addressed by instructors who desire to effectively prepare their students. This action research study reports on the attempt to incorporate a constellation of argument-related activities into an online AP Physics 1 course in order to investigate the role they may play in argument skill development.

The study was conducted with homeschool students who took AP Physics 1 through an online education company called Physics Prep. Students were exposed to a variety of argument-related activities as they learned physics content over a six-month section of a school year. Analysis of data produced in pre/posttests, pre/poststudy interviews, and student-constructed artifacts, allowed for answers to inquiry questions to emerge over time. These questions were associated with the precourse level and subsequent development of skill in argumentation for fifteen students.

Findings indicated that the incoming skill level of students was generally in need of improvement relative to the expectations set by the College Board. Change in argument-related performance for course activities was then measured over time. Group-level improvement was generally found. However, inconsistent individual exhibition of argumentation skill was also noted, which pointed to sufficient physics-related conceptual development as a necessary foundation on which arguments in AP Physics 1 can be constructed. Without conceptual grounding, the arguments evaluated in this study lacked the quality required by the College Board even when the structural understanding of a high-quality argument was evident in other work. A rationale is offered that may explain the connection between argument construction and conceptual understanding of physics subject matter.

Table of Contents

List of Tables

List of Figures

Preface

Although counter-intuitive, it is often said that the longer one works in any profession, the clearer it is how much one doesn't know. This is definitely true for me. The sense of appreciation I have for others who supply missing knowledge has deepened with age. Certainly, the learning community at the University of Pittsburgh's School of Education, faculty and students alike, were important sources of insight for me as I navigated the EdD program as a nearly sixty-year old veteran teacher. I am particularly grateful for my advisor, Dr. Ellen Ansell. Rarely have I encountered such a high-level of expertise combined with humble strength and clear focus in a colleague. Ellen, thank you for your patience and wisdom. I extend the same thanks to the other members of my committee, Dr. Andy Cavagnetto and Dr. Kari Kokka. Each of you has uniquely influenced the way I view the role of science education in society.

There is only one thing in my life more important to me than learning. It's the relationships I have with my family. My wife, Anne, has supplied such love and support to me for nearly forty years that when she or I struggle, we both struggle. When either of us celebrate, we both celebrate. Never has that dynamic been more apparent as during my time in the EdD program, particularly when she helped me to analyze the qualitative interview data! Our bond is special, and it means everything to me. I also want to acknowledge the support of my children in everything I do, including my work at Pitt. Sarah, Elizabeth, and Jackson have been the most amazing sources of inspiration in my life. Finally, I want to thank my twin brother, Mark Kernion, and my friend Colleen Schmiech, for sharing their expertise and their willingness to help in the analysis of the data I collected for this study.

1.0 Introduction

"If we hope to overcome education's preoccupation with standardized test scores, we need something better to replace them—a richer vision of what we want education to accomplish."

Deanna Kuhn (Kuhn & Shaughnessy, 2010, p. 271)

During my more than three-decade career as a science educator, I have been searching for an ever-richer vision of what can be accomplished in the classroom. Although I've been satisfied by the yearly evolution of new students from novice learners to various levels of confident problem-solvers, I've also been humbled by the annual realization that I can get better at the practice of teaching. I believe that educators are motivated to improve because they have a profound internal understanding that when they become better teachers, the benefit of that development is multiplied through their students, who then pass it on to others. It's a dynamic not available in most professions.

Over the years, I've tried to adapt my teaching to become less teacher-centered and more practice-based. I haven't always been successful, but I've attempted to learn from what wasn't effective as much as from what worked. Although I spent all but the last four school years teaching physics in a traditional suburban high school, and currently teach AP Physics to online students through a company I founded called Physics Prep, I've found that the challenge of yearly improvement to be consistent in both settings. Of all the strategies, techniques, insights, and technological gadgets I've incorporated into my pedagogical repertoire, nothing has as much promise, in my opinion, as simply and conscientiously promoting the "grasp of practice" in my students (Ford & Foreman, 2006). That phrase implies that science students can be considered successful when they internalize and practice what it is that real-world scientists actually do. This is not to diminish the status conferred to students who also learn what scientists know. But knowing is not the same as doing. Ultimately, I want my students to be labeled successful from both perspectives.

When it comes to science practice, nothing is more basic than arguing. Of course, this isn't the type of arguing that involves yelling and mean-spiritedness. Instead, it is an integral part of a community-based search for deeper understanding of the world around us. Science practice is a cooperative venture that is one of the most exciting available to humans. To be part of that adventure requires certain skills, one of which is the ability to argue from justified evidence. The investigation outlined in this report describes my 2018-2019 attempt to foster the development of that skill in my students. My 2019-2020 attempt at improvement will be informed by it.

1.1 Problem Area

The primary problem area addressed in this investigation is the general challenge associated with the development of evidence-based argumentation skill in science students. The importance of this topic is anchored in wide-spread agreement within the educational research community that acquisition of argumentation skill is central to the attainment of scientific literacy (Duschl, 2008; National Research Council, 2012; Asterhan & Schwartz, 2016). Kuhn (2010, p. 810) states that "A conception of *science as argument* [emphasis added] has come to be widely advocated as a frame for science education." The inference is that when science teachers neglect to address argumentation as part of instruction, they miss an important opportunity for student learning. However, there is a lack of clarity from educational researchers on effective instructional strategies that should be employed to develop argumentation skill (Cavagnetto, 2010).

Importantly, Lawson (2004) claims that we do understand *how* students develop the intellectual skill needed for students to successfully argue in a scientific way. He says that teachers, "…must teach in ways that allow students to develop the necessary reasoning abilities. In short, instruction must not only "fit" students' current developmental levels, but it must also provoke students to progress to higher levels. The evidence is clear that the best way to do this is to teach science in the way science is practiced (Lawson, 2004, p. 333-334)." Although that idea is easy to agree with, identifying the manner in which it is best accomplished is another matter altogether. Each science course and classroom have their own cultures, personalities, and expectations. The instructional strategies needed to promote the development of skill in argumentation may differ in each case, but the importance of teachers attempting to find out what works for their students is a consensus opinion within the research community.

According to Zohar and Nemet (2002, p. 38), an argument can be defined in several ways, but generally "… consists of either assertions or conclusions and of their justifications, or of reasons or supports." A commonly used argument model introduced by Toulmin (1958) identifies the three components of claim, evidence, and justification as the *minimum structure* needed for creation of a valid argument. The terms argument and argumentation are distinguished by the idea that an argument is a product of the process of argumentation (Kuhn & Udell, 2003; Osborne, Erduran, & Simon, 2004). Argumentative discourse is the act through which an argument is produced. When involved in evidence-based argumentation, students must critically evaluate their own claims and assess the claims of others. Development of argumentation skill not only improves discipline-specific conceptual understanding and occupational preparedness, but also promotes the civic duty of every citizen to effectively participate in public debate on scientific issues (Sandoval, 2005). Nevertheless, the science classroom, as it is traditionally structured, is rarely a space that supports learning how to argue scientifically (Kuhn, Hemberger, & Khait, 2016; Osborne, Erduran, & Simon, 2004). Use of scientific argumentation as a teaching strategy does not fall within the standard repertoire of instruction found in typical science classrooms (Jimenez-Aleixandre, Rodriguez, & Duschl, 2008). Osborne (2009) claims that science teachers do not commonly have the training needed to effectively engage students in using argumentation-to-learn instructional techniques. Hence, the primary challenge facing practicing teachers who want to use argumentation for learning has two features. The first is the lack of agreement on *how* to use argument as an effective learning tool and the second is *lack of teacher training* for those techniques that have shown some promise.

As this inquiry focused on argumentation in an online setting, a secondary problem area was also pertinent. It is the challenge encountered by online science teachers when social interactions are important for learning. Crippen, Archambault, and Kern (2013) claim that online science instructors confront several inadequately investigated issues not found in traditional education. These include how to effectively address the social aspects of science practice and assess progress in learning science skills when students work in isolated settings. Activities such as cooperative lab investigations and engagement in argumentation are good examples that are sometimes difficult to accommodate for in an online course.

Taken together, the two problem areas identified above combine to create the overall backdrop for this inquiry: *Research shows that learning to scientifically argue likely benefits students. However, there is disagreement regarding the most effective instructional strategies that should be employed to promote the development of scientific argumentation skills.*

1.2 Problem of Practice

Current research suggests that learning by K-12 science students should be assessed for more than understanding and application of discipline-specific content via computationally-based problem solving (National Research Council, 2002; National Research Council, 2012). For a student to be deemed scientifically proficient, she must also be able to voice explanations of natural phenomena, communicate her understanding of scientific principles, and create, revise, rebut, and refine scientific arguments (NGSS, 2013; Reiser, Berland, & Kenyon, 2012).

This research-based understanding has prompted the College Board to revise the curriculum framework for its Advanced Placement science courses in Biology, Chemistry, and Physics (Drew, 2011; Magrogan, 2014). Two algebra-based AP Physics courses (AP Physics 1 and AP Physics 2) were recently introduced, each with a clear emphasis on the importance of student understanding and demonstration of science process skills (College Board, 2012). The exams associated with these courses also reflect the latest educational research, expecting students to express their understanding of physics in a wide variety of ways. In addition to answering multiple choice questions, students who take the AP Physics 1 exams must respond to open-ended questions that require outlining an experimental design, translating qualitative and quantitative data, and writing paragraph-long, well-structured arguments in defense of a scientific claim. It is the required proficiency in argumentation that specifically motivates this inquiry. Stated plainly, I am deeply interested in knowing how I should support students in meeting this demand.

I believe that the instructional strategies and classroom culture required to prepare students for these challenges differ from those needed when the algebra-based, computationally-heavy, AP Physics B course was offered prior to the 2014-2015 school year. Strategies suggested by the College Board (2012) include extensive use of sufficiently-open, inquiry-based lab activities, encouragement of scientific argumentation among peers, and the demand for evidence-based student writing that exhibits effective communication and deep understanding of concepts. These strategies represent a change and a challenge for AP Physics instructors (Fullerton, 2017). The importance of needing to address this challenge is made clear when one analyzes the data on algebra-based AP Physics test scores over the previous several years (see Table 1). The decline in scores is likely due to several factors, one of which (I hypothesize) is the requirement that students must exhibit skill in scientific argumentation that is not commonly addressed in the classroom.

Table 1. Algebra-based AP Physics scores 2013 to 2016 (College Board, 2019)

Score	Physics B 2013	Physics B 2014	Physics 1 2015	Physics 1 2016
5 (extremely well qualified)	16.6%	15.8%	5.0%	4.6%
4 (well qualified)	19.9%	18.5%	13.6%	14.0%
3 (qualified)	26.1%	26.5%	20.7%	21.2%
2 (possibly qualified)	16.3%	17.0%	29.8%	30.2%
1 (no recommendation)	21.1%	22.3%	31.0%	30.0%
Number of tests administered	89263	93574	171074	169304
Mean Score	2.95	2.89	2.32	2.33

Comments published by the College Board, written by AP Physics 1 exam graders as advice to teachers, consistently mention a lack of polished argumentation skill seen in student responses to assessment items that require them (College Board, 2015a, 2016a, 2017a). This indicates that something is steadily missing from instruction that needs to be addressed. Although I do not know if that’s true for my own students, as I only have access to their overall AP scores, I sense the need for a greater understanding of the general level of argumentation skill that I help my students attain.

Facilitating online courses at Physics Prep, with students who are not physically present to each other, creates a challenge, specifically for lab investigations and the development of argumentation skills. Effective use of ever-changing technology to implement alternatives to strategies that may be effective in brick and mortar schools is not always as easy or straightforward as it might seem. This is particularly true when students are not associated with a specific institution that may supply common resources to its students, as is the case for homeschoolers at Physics Prep. The inquiry described in this report addressed a concern I had regarding the development of argumentation skills for students who take the online AP Physics 1 course at Physics Prep: *Are my students sufficiently prepared to respond to questions that demand skill in evidence-based argumentation?*

1.3 Inquiry Questions

This investigation was designed to address three questions whose answers provided stepping stones for progress in my teaching practice. Like all good queries, they led to other questions as part of an ongoing effort of improvement. The questions addressed in this study are ones that have been on my mind for several years. Although they are specific to my practice, I am confident that similar ones are being asked by science teachers of all levels and disciplines. They reflect a contemporary concern that is congruent with the societal demands our students must face as they move beyond the classroom. In short, are we effectively preparing our students for participation in an increasingly scientific and technical world?

Inquiry Question 1

What is the incoming level of argumentation skill of my students?

Inquiry Question 2

What patterns of progress in argumentation skill development do students exhibit as they engage in the constellation of argumentation activities offered at Physics Prep?

Inquiry Question 3

Are the resources offered at Physics Prep sufficiently robust to provide students the opportunity to reach levels of argumentation skill needed to effectively respond to assessment items they will encounter on the AP Physics 1 examination?

2.0 Review of Literature

In her introduction to *Perspectives on Scientific Argumentation*, Deanna Kuhn (2012) discusses the surge of interest in recent educational research devoted to argumentation. The ubiquity of articles found during this literature review has verified her assertion. Because of this, a review of reasonable length will inevitably present only a partial perspective that may neglect some important findings on the subject. Although not universally true, the studies discussed in this review are frequently cited in scholarly articles and represent research that skews somewhat toward argumentation skill development that takes advantage of new technologies such as web-based instruction.

2.1 General Findings

My interest in this topic is grounded on the claim that developed argumentation skills allow for the construction of better arguments and for more effective argumentative discourse (Kuhn & Udell, 2003). How to best develop those skills, however, is still an open question for the research community (Kuhn, 2012).

There are four general findings I will discuss that arise from the literature. These are:

1. Use of argumentation skill is related to epistemic commitments.
2. The effectiveness of various instructional strategies designed to develop argumentation skill has not been conclusively determined. Findings show mixed results.

3. The identification of the goal of argumentation skill development is important when designing appropriate instructional strategies and learning environments.

4. Assessment of argument quality is not an easy task in either online or traditional classes.

2.1.1 Epistemic Aspects of Argumentation

Disposition toward and competence in using epistemological strategies like argumentation work together to motivate their use (Kuhn, 2001; Perkins & Tishman, 1993). Inclination to pursue knowledge is partly determined by the value one places on that pursuit. Is there a point to using the strategy? Is it worthwhile? Does one possess the necessary skills? Kuhn (2001, p. 8) summarizes this concept stating, "People must want to know, and appreciate the benefits it confers, if they are to undertake the effort it requires." She claims that the epistemic characteristics that apply to arguments in general, like internalizing the value of argumentation, cut across disciplines and can be viewed as important in and of itself.

Kuhn has long been an important voice in the academic discussion about the importance of learning argumentation skills. Her efforts to understand the *why* behind *what* people think (Kuhn, 1991) has pushed her to recognize the importance of learning argumentation skill as a fundamental part of education (Kuhn & Shaughnessy, 2004). Kuhn (2010) has expressed concern that the development of ambitious content knowledge simultaneously with argumentation skill development may be difficult for students. However, the practice of this skill within a content-grounded context is important and should be done across diverse settings to be reinforced. A recurrent theme in her work is that intellectual skills stand above subject-specific contexts, but that context is by no means irrelevant.

Kuhn (2001) describes the progression through levels of epistemic understanding as movement from *absolutist*, to *multivist*, to *evaluativist*. In short, the evolution proceeds from an understanding of a single perspective, to the appreciation of all perspectives where no evaluation is possible, and finally to the understanding that although alternatives may be legitimate, the strength of an argument for a perspective is worth evaluating. This evolution varies among the population, giving rise to differences in how people cognitively perform.

Kuhn (2010) believes that the epistemic aspects of argument cannot be simply told, but must be appropriated through student practice. In other words, trusting that science can be a reliable source of knowledge is acquired experientially. Kuhn (2010) describes a research project with middle schoolers wherein teams of students investigate a *non-scientific* topic, engage in electronic dialogues with other students, strategize for a final debate, and then participate in the debate itself. The students are then shown video of the debate as a reflective activity. She reports that progress initially occurs when the realization of attending to the claims of another is made by a student. That understanding moves the argument from simple claim-making to claim critique. This is followed by the recognition of the need for evidence, both for the support of a claim and, less commonly, to critique the claim of another. Importantly, she also states that evidence is not just used, but is also discussed, implying metalevel understanding of its importance.

Kuhn (2010) also describes research which tests the same methodology for *scientific* topics. Her findings indicate that gains in argumentation skill occur with science topics as well as with non-scientific topics. However, the transfer of learned skills from science to non-science topics is more pronounced than from non-science to science topics. She uses these findings to justify the use of argument strategies in the science classroom and comments on the nature of science as the reason that skills are unevenly transferred. Kuhn (2010) believes that science discourse does not exhibit the relativism that might be acceptable in socioscientific discourse. The challenge in the socioscientific context is to constrain the view that “anything goes”, whereas that of the scientific context is to recognize that there are social aspects to science which give rise to different, but constrained, perspectives. Similarly, Osborne, Erduran, and Simon (2004) claim that development of argumentation skills in a scientific context is simply more difficult than in a socioscientific one, due to the knowledge base required to present evidence and justifications.

2.1.2 Instructional Strategies

Cavagnetto (2010) wrote an extensive overview of research findings for more than fifty studies related to instruction intended to develop argumentation skill in science students. He found that instructional strategies fall into three general categories. These include immersion (learning by doing), direct (explicit) instruction of argument structure, and discussion of socioscientific topics. I will first discuss research in support of direct instruction (approach 1) and then turn my attention to that which supports immersion (approach 2). Comments on socioscientific strategies will be a part of each discussion.

Although the categories listed above are helpful when discussing these strategies, the boundaries between them are not always clear. Proponents of each approach use common educational terms such as scaffolding, epistemic practice, context, metacognition, and many others as descriptive of that approach. Claims are often nuanced and can be confusing to an educator in search of best practice. As has been found for other supported-by-research instructional strategies such as inquiry-based teaching, conversion from research to practice can be difficult, and when attempted can be ineffective for many reasons (Cavagnetto and Hand, 2012). Utilizing instructional strategies intended to bolster argumentation skills is no different. Berland and Reiser

(2009, p.28) acknowledge the two approaches described above, but also recognize their common goals, stating “Their differences lie in the aspects that they choose to emphasize and how it is made explicit through their interventions.” It’s fair to say that regardless of which approach is in the foreground of a particular study, the other lingers in the background or is found in plain sight. For this reason, the following sections, while attempting to highlight the benefits of each approach, will necessarily repeat the common theme that it is difficult to fully disentangle them.

2.1.3 Direct Instruction

Explicit instruction of argumentation directly and purposefully identifies the structural features of an argument for students to learn. These include components such as claim, evidence, and justification (rationale) statements. It is hoped that through exposure to definitions and examples of each component, students will be able to apply this knowledge to create their own arguments and evaluate those of others. Although research that supports the value of explicitly teaching argumentation is plentiful, all pertinent studies also include student engagement in the process *of* argumentation. I present details of two studies here, but others report similar findings (Erduran & Aleixandre, 2008; Larson, Britt, & Kurby, 2009; McNeill, 2009; McNeill & Krajcik, 2008). Zohar and Nemet (2002) investigated the effect that explicit instruction of reasoning skills had on learning in a ninth-grade socioscientific genetics unit. They claim that “…general and specialized knowledge function in a strong partnership” (Zohar & Nemet, 2002, p. 37), implying that general argumentation skill cannot be separated from the context in which it is practiced. Principally, instead of worrying if patterns of thinking are generalized or content-dependent, one should recognize that they work together, although perhaps differently, in any given circumstance. A hybridization of direct instruction and immersion strategies is, in this sense, implied whenever research is done on the effectiveness of the explicit instruction of argumentation. Through analysis of both audio transcripts and written work, Zohar and Nemet (2002) found gains in argument quality and content knowledge for students exposed to explicit instruction of argumentation. The improvements were not found in a comparison group. These gains, they claim, may result from metacognitive activity prompted by student reflection on reasoning and on shifts in classroom culture wherein the value of argumentation is clarified and upheld.

A second oft-cited example of the attempt to use explicit instruction of argumentative structure is work done by Osborne, Erduran, and Simon (2004) with junior high students in England. Scaffolded argument construction-aids called writing frames were used, some including sentence stems, that helped guide student work. Additionally, evaluation of argument strength was introduced to students through discussion of exemplars designed to model good argumentative practice. Classroom data (video and audio transcripts, field notes, and interviews) were collected at the start and the end of an intervention that involved teaching argument in a scientific context over the course of the school year. The data show that an *increase in the amount of argumentative discourse* is possible when instructors focus on both epistemic and social considerations. The authors also analyzed oppositional episodes found in the transcripts using a five-level framework designed to judge argument quality (Table 2).

Table 2. Levels of argument quality used by Osborne, Erduran, & Simon (2004)

Quality Level	Description
1	Argumentation consists of arguments that are a simple claim versus a counterclaim or a claim versus claim.
2	Argumentation has arguments consisting of claims with either data, warrants, or backings, but do not contain any rebuttals.
3	Argumentation has arguments with a series of claims or counterclaims with either data, warrants, or backings with the occasional weak rebuttal.
4	Argumentation shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counterclaims as well, but this is not necessary.
5	Argumentation displays an extended argument with more than one rebuttal.

Although the most common level of argument was at level 2, both at the beginning and the end of the year, the number of higher-level arguments (levels 3 and above) increased while the number of level 1 arguments decreased. Additionally, the authors found that socioscientific contexts produced a greater percentage of higher-level arguments than those done in a scientific context, but there were no significant differences in the improvement found between the experimental and comparison groups in the study for either context. Osborne, Erduran, and Simon (2004) conclude that although the improvements for the experimental group were smaller than anticipated, changes in the attitudes of teachers toward the use of the instructional strategies promoted in the study were noteworthy and positive. Their overall claim is "…that improvement at argumentation is possible if it is specifically addressed and taught" (Osborne, Erduran, and Simon, 2004, p. 1015), but that more than a nine-month intervention may be needed.

2.1.4 Instruction Through Immersion

Similar to learning a new language, proponents of immersion strategies for learning argumentation claim that appropriation of this skill is accomplished through active participation in its practice (Cavagnetto, 2010; Manz, 2015). A succinct way to frame that main point is generalized by Rapanta, Garcia-Mila, and Gilabert (2013) by claiming that arguing to learn (immersion) is different and better than simply learning to argue (explicit instruction). Proponents of immersion strategies voice concern that practice can become confused with structure when the latter is brought to the foreground as it is with explicit instruction (Manz, 2015). For example, Berland and Hammer (2012) do not view the difficulties students encounter in learning these skills as related to something missing in their conceptual development as is implied with explicit instruction. They cite recent work that indicates the presence of argument skill, albeit underdeveloped, found in young students. They claim that when student conceptual development of *how to argue* takes precedence over the understanding of *what is happening in the classroom with respect to knowledge*, there is a risk of confusion between argumentation and pseudo-argumentation. The latter is motivated by a benefit not associated with the intended epistemological goal of the instruction. In pseudo-argumentative discourse, students may create an argument structured to please a teacher rather than to learn something. Berland and Hammer (2012) state that explicit instruction should only be utilized to solve a problem indicated by a struggling student. Argumentative activity must precede direct instruction and be responsive to that struggle.

The foundation of the immersion strategy, as applied to science practice in this case but also to learning in general, lies in sociocultural theory whose origins are based in the work of Russian psychologists Vygotsky and Bahktin (Ford & Forman, 2006). The general idea is that a student appropriates knowledge through activities rooted in a social and cultural context. Some theorists place knowledge as situated *primarily* in the sociocultural context, and *secondarily* within the individual acting in that context (Hickey, 2015). Research that supports immersion strategies is easy to find. Although the sociocultural aspects of the interventions included in this category are the focus of these studies, many do rely on extensive use of scaffolds which aid students in the construction and evaluation of arguments. While not necessarily explicitly instructing a student about argument structure, a scaffold introduces technical terms and guides a student along a path. Both guided writing (Berland & Reiser, 2009) and computerized aids (Bell & Linn, 2000; Iordanou & Constantinou, 2015) have been used to scaffold the immersion of students in argumentative activities.

A typical example of this approach can be found in work by Sandoval and Reiser (2004), who made use of an electronic journal called ExplanationConstructor as a way to organize student investigations. ExplanationConstructor provides discipline-specific scaffolds designed to help students both construct and evaluate explanations. This tool helps students make distinctions between claims and evidence and link them together in an organized product that can be used reflectively. As a way of clarifying their overall approach, Sandoval and Reiser (2004) state “We cannot overemphasize the importance of the fact that students’ use of ExplanationConstructor and other tools in this curriculum was embedded within tasks that were specifically aligned with the epistemic practices we are trying to develop. This point may seem obvious, but we stress it because it is crucial to understand that students’ performances here do not originate from the tools but are supported by them.” Sandoval and Reiser (2004) do not want their work to be confused with support for explicit instruction of argumentation. They carefully distinguish between conceptual tools that aid student reasoning from epistemic tools that assist articulation of understanding of phenomena. Interestingly, Sandoval and Reiser (2004) also discuss that when updates to rubrics and adaptations to software used in their research more explicitly link components of argument structure, student argumentation improved. One wonders if the tools they describe as enablers of classroom norms didn’t also serve as explicit instructors. In any case, the findings presented in their work indicate that epistemic tools such as ExplanationConstructor can be effective in supporting student inquiry.

As a final example of an intervention that supports immersion strategy, the work of Ryu and Sandoval (2012) with elementary-level science students stands out. They conclude that an *orientation to argument* as a discursive classroom strategy can develop epistemic awareness of science argumentation skills in students. In other words, when students are expected to participate in argumentation and value the benefits it brings, they are more likely to engage in the practice. They suggest that rather than being viewed as a supplementary skill, scientific argumentation should be understood as part of the culture of the classroom in which students are immersed. This insight is supported by Berland and McNeill (2010) when they claim that "Developing a classroom culture and norms is also essential for supporting student engagement in the argumentative process…teachers can facilitate the creation of these norms by creating situations in which it makes sense for students to engage with one another's ideas." Ryu and Sandoval (2012) warn that this practice must occur over a significant amount of time, claiming that research on interventions lasting short periods of time have shown mixed results.

2.1.5 Hybridized Instruction

When discussing immersion strategies vis-à-vis scientific argumentation skill, it is helpful to refer to immersion approaches used in second language learning. Such learning occurs when a student is immersed in second language practice without an emphasis on learning grammar. This instructional strategy has many variations, but is fundamentally different from the type of immersion used for first language acquisition. Genesee (1985, p. 543) describes immersion as "…not simply a matter of treating second language learners as if they were native speakers of the target language…The effectiveness of immersion depends very much on the quality of the interaction between the teacher and the learner." In other words, immersion in a second-language classroom is not the same as immersion in the real world, where an individual is plopped into a new environment and learns how to get along on her own through raw experience. Similarly, immersion in classroom argumentation is not the same as actual science practice, and its effectiveness is contingent upon many classroom features. Accordingly, I will describe findings that corroborate the belief that explicit instruction of the structural aspects of an argument combined with scaffolded strategies connected to immersion activities may produce the greatest benefits for the development of argumentation skills.

Kuhn and Udell (2003) describe an intervention whose results indicate that argumentative discourse, while vital to skill development, is not sufficient to account for the gains in argument quality and effectiveness that they measured. Structured scaffolding of the process is necessary as a means of representing concepts externally. McDonald and McRobbie (2010) produced a comprehensive overview of educational research on the question of whether explicit instruction or immersion is best practice for improving student understanding of the nature of science and for developing argumentation skills. Their conclusion is that explicit instruction (structure, function, application, assessment) for *scientific* argumentation is needed for students to improve their skills. They suggest, however, that this may not be true for *socioscientific* argumentation. Thus, the context of the instruction is important. Additionally, they think it crucial for students to *participate* in argumentation. They suggest that one's views on the nature of science will affect one's willingness to engage in that practice and cannot be replaced by structural knowledge of argumentation in and of itself. Thus, epistemic orientations (absolutist, multivist, and evaluativist), as described by Kuhn (1991) that affect the attitude toward and the frequency of engagement in argumentation, play a role in the process of appropriating these skills. Clearly, something complex is going on when students engage in argumentation (Berland & Reiser, 2011) and teasing out the effects of various instructional approaches in not easy. Perhaps a hybridized approach to instruction of argumentation covers all the bases and may be essential for all student learning needs to be met.

Citing recent research, Manz (2015, p. 561) states that, “Making argumentation structures visible to students encourages them to make their ideas explicit, promoting the elaboration, connection, and consolidation of scientific understanding…In spite of the way that these studies have illuminated how students tend to think about argumentation and shown how rare it is in classrooms without explicit support, there have also been critiques of these approaches, on the grounds that they are not sufficient for allowing students adequate access to the activity systems within which scientific argumentation operates.” Cavagnetto and Hand (2012, p.43), when discussing the spectrum of practices designed to develop argumentation skill, emphasize the importance of adopting a “middle ground” but also warn that explicit instruction of structure can promote a false sense of independence between claim making, evidence finding, and reasoning used to justify the argument.

Any thorough review of the literature will undoubtedly leave the impression that scientific argumentation is a complicated skill to develop. As a way to demonstrate the complexity of argumentation, Berland and Hammer (2012) discuss the concept of framing, or determination of *what’s happening,* in relation to argumentative practice. Frames are schemas used by individuals to contextualize the moment. They claim that student engagement is prompted by framing through *seeing the point* of a classroom activity. Frames organize past experiences and interpret present experience. Because the present moment provides shifting clues as to what is going on, framing is dynamic. One moment a student may frame the lesson as open inquiry, with no pressure to determine anything specific, which then changes suddenly to the frame of *find the correct answer*. The clues students receive about *what is going on* are understood through both overt discourse and metacommunication. Because framing, by its nature, is variable, student understanding should be expected to show variability.

Even though the complexity of argument skill development and the various approaches suggested to support it can be confusing and appear polarizing, there are certain aspects of argumentation that establish common ground among proponents of both explicit instruction and immersion strategies. One such dimension is the positive role played by student reflection. Like Zohar and Nemet (2002) who investigated direct instruction, Iordanou and Constantinou (2015) emphasized the contribution that reflective activity has in creating high-quality arguments through immersion. They based their work on the concept that students struggle with argument construction due to lack of metalevel understanding that is both strategic and epistemological. Since their intervention was grounded in student dialogic argumentation conducted online via instant messaging, a transcript of the process was available for student reflection. The reflection activity included using a worksheet that asked students about their use of evidence and justification *after* the dialogic activity had taken place for purposes of revision. Unlike Zohar and Nemet (2002), this intervention did not focus on scaffolded individual argument construction, but instead on scaffolded dialogic argument analysis. Participants in both the intervention group and the comparison group were evaluated regarding argument skill before, during, and after the intervention. The findings, like those of Zohar and Nemet (2002), indicate that the intervention group showed increases in the use of evidence whereas the comparison group did not.

Kuhn and Shaughnessy (2004) make a similar point, noting that the development of argumentation skills depends upon meta-level shifts in strategy rather than improvements in the execution of already-developed strategies. Kuhn and Udell (2003) claim that such change can be prompted by dense exercise of argumentation skill and that the exercise must be goal-oriented and give opportunity for performance reflection. Such reflective-driven shifts are not limited to high-achieving students. They have been found in lower-achieving students as well (Yerrick, 2000).

It seems reasonable to conclude, based on the general body of literature reviewed, that a hybridized approach to instruction may be helpful in the appropriation of argumentation skills (Cavagnetto, 2010; Kuhn, 2012). Including both explicit instruction and immersion activities for students may provide a best-of-both-worlds scenario *if* one can avoid the pseudo-argumentation concerns expressed earlier. In other words, if explicit instruction of argument structure takes on too strong a role in lesson design, students may miss the larger point by framing the experience as content independent. This could make the educational experience lack the science learning it was meant to instill.

2.1.6 Persuasion, Articulation, and Sensemaking

Regardless of which instructional strategy is chosen to facilitate the learning of argumentation, Cavagnetto and Hand (2012, p.39) claim that "Like inquiry, the effectiveness of argument is dependent on the goal of instruction." Berland and Reiser (2009) identify three goals associated with argumentative practice that help one understand how explicit instruction and immersion work together. These are sensemaking, articulating, and persuasion. Sensemaking requires discipline-specific knowledge and relies on the understanding of content. Articulation demands skill in communication and rhetoric. Persuasion requires the argumentative practice have a social aspect connecting sensemaking to the learning community in which the process is taking place. Each of these goals are intertwined but also unique, particularly with respect to the level of their development found in an individual student. Highly developed sensemaking, for example, does not guarantee that an argument will be persuasive. Each of these goals identify a different aspect of argumentation that should be addressed during instruction through strategies and supports that are not necessarily the same but sometimes can be. For example, the goal of sensemaking may be supported by teaching how to articulate argument components through direct instruction. Alternately, sensemaking can be developed in parallel with persuasion if learning is assumed to be appropriated socioculturally (Berland and Reiser, 2009) within an immersion activity.

Thomas Kuhn's (1970) work, *The Structure of Scientific Revolutions*, initially published in 1962, was among the first to acknowledge the role the scientific community plays in establishing truth about the physical world. Sociocultural concepts such as those proposed by Kuhn and others point to scientific knowledge as being communally constructed (Newton, Driver, & Osborne, 1999). These ideas have deep and lasting influence on how philosophers of science and other academics view the nature of science (Sandoval & Millwood, 2015). The basic roles that scientists play during their social practice are those of creators and critiquers of claims (Ford and Forman, 2006). The interplay in which scientists engage, between representing the natural world through inherently uncertain claim-making and subsequent argumentation about those representations, is a fundamental aspect of science practice (Manz, 2015). Thus, it's not surprising that persuasion is viewed as the ultimate goal of scientific argumentation (Berland & Reiser, 2009; Sandoval & Millwood, 2015), and why persuasion rather than performance may be the best frame through which students view argumentative practice (Berland & Hammer, 2012).

Berland and Reiser (2009) suggest that having persuasion as a goal benefits argument construction. They claim that there are two strategies a student can pursue in order to construct a written argument (of the type typically required on an AP Physics 1 exam). The first is when the reasoning made in the argument is intertwined with evidence. Although the argument may not contain incorrect elements, this construction can be confusing to the reader. The second is when the two are carefully distinguished so that a reader can easily identify them in the explanation.

Their research indicates that when the goal of persuasion is pursued, students are more likely to construct arguments of the second type. McNeill and Krajcik (2008) support this idea conversely, claiming that if explicit instruction on the components of an argument is done *without emphasizing the purpose* (goal) of the practice, argument construction becomes algorithmic and ineffective. Typical school science practice can be differentiated from authentic science practice for this reason, as sensemaking often assumes the lead role in school science whereas persuasion is paramount in actual science practice. Jimenez-Aleixandre, Rodriguez, & Duschl (2000, p. 759) use the phrase "doing the lesson" as opposed to "doing science" as a way to capture this distinction. They claim that to reason in a scientific way means having to defend the choices one makes though argumentation and that students should engage in persuasive argumentation, where the goal is to convince someone of your claim.

Alternatively, Manz (2015) discusses the importance of the distinction between published arguments made by scientists that are meant to persuade and the collaborative arguments made by practitioners along the way towards that goal. The former emphasizes disputative argumentation while the latter focuses on deliberative discussion. Persuasion is the primary goal in disputation. Arguments that are disputative defend claims that the scientist wants the world to see and is meant to have influence. But if one forgets about the process through which a published scientific article is based, deliberative argumentation, one "erases the historical and personal activity that supports it (Manz, 2015). Deliberative argumentation is more about sensemaking and may be the appropriate primary goal of school science. Manz (2015, p.569) promotes the development of a student "activity system" whose context provides students the opportunity to engage in deliberative discussions whose norms are clearly understood and where "productive uncertainty" can be resolved. Work currently underway by Andy Cavagnetto at Washington State University is investigating the benefit of deliberative discussion in large biology course lectures (Cavagnetto, 2018, personal conversation).

All of these insights can play a role when instructors plan lessons meant to develop argumentation skill. The crucial question to ask is, “What is the goal of the instruction?” If sensemaking is primary, the instructional choices may be different than when persuasion is the top goal.

2.1.7 Affordances and Constraints in an Online Setting

Some of the issues discussed above point to possible problems that may arise when teaching scientific argumentation to online learners. These generally revolve around the physically isolated learning space in which online students participate in a course. However, the new affordances made possible through web-based education tools also provide opportunities not seen before in the history of education. For example, online video meeting applications allow for students and teachers to meet and share work in a virtual space in much the same way they would in a classroom. Unlike other emergent technologies of the past, there is reason to believe that the cyberinfrastructure of today consists of the social, cognitive, and technological features that may allow for revolutionary change in education (Martinez, & Peters Burton, 2011). Instead of seeing the rapid increase of online learning as making the educational endeavor more difficult, it may be best to accept that “…the world has developed in such diverse directions and created new and particularly complex demands for citizenship, college, and careers that it is no longer possible for old learning environments associated with old learning paradigms to accommodate them.” (Avgerinou, M., Gialamas, S., & Tsoukia, L., 2014, p. 329).

Benson (2003) lists several features of online learning as affordances not often found in a traditional classroom environment. These include the ability of online instructors to engage all students equally, and the possibility of immediate feedback. Among the strategies she suggests online instructors use in a purposeful way are ones that relate to the development of student argumentation skills. For instance, carefully designed virtual discussions are effective when developing skills that require the use of high-level cognitive skills, such as those needed when constructing and critiquing arguments. Others include peer assessment and self-assessment that Benson (2003) claims are easy to effectively facilitate in an online setting, as long as well-understood, rubric-based methods are used in the assessment process.

The common absence of face-to-face interactivity among students and between students and the instructor, along with physical isolation during learning, have both been raised as constraints associated with online learning (Sher, 2009). Sher's (2009) research suggests that these concerns can be alleviated by creating online learning environments that purposefully address the social nature of learning. For example, Pifarré et al. (2014) report positive results using a web-based application called Metafora, wherein visual tools promote what they term *Learning to Learn Together* (L2L2) via virtual discussion and argument building. They contrast L2L2 with *Learning to Learn* (L2L) strategies in which individuals learn in isolation. Metafora's visual tools are accessed through linked-together task icons that promote the management of problem-solving within a social framework such that learning is viewed as a communal endeavor. The context of the learning is created through social interactions that are reported to be crucial to student progress.

There are several types of interactions than can occur within an online community of learners. These include student-teacher, student-student, and student-content types (Moore, 1989). However, Garrison and Cleveland-Innes (2005) assert that simply quantifying the amount of interaction is not a good indication of meaningful online learning. They suggest effective online learning must be purposeful and systematic. Furthermore, they claim that an important distinction exists between simple online interaction and what they refer to as online *presence*. Student presence can be categorized as having social and cognitive dimensions. Garrison and Cleveland-Innes (2005) report that social presence in a community of learners is often developed through student-student interaction, whereas the cognitive dimension is primarily established through student-teacher relations. Teacher presence implies more than basic discourse and direct instruction with a student. Although depth of student learning is influenced by the learning environment *structured* by the teacher, instruction that promotes teacher presence is not teacher-centered. Instead, the teacher is a moderator, setting clear, achievable goals that are consistent with desired learner outcomes and creating a well-designed and organized learning space. Both student presence and teacher presence are indicative of a high-level of interactional quality in the online setting.

Clark et al., (2007) outline various features associated with the online learning environment that may be superior to traditional settings when promoting skill development in argumentation. These include:

1) Collaborative communication (synchronous and asynchronous).

2) Co-creation and sharing of artifacts, such as in-depth concept maps and other intellectual documents, compared and refined over time.

3) Enriched access to information through extensive online sources, including data and visualizations, that are a part of software programs designed to help foster argumentation skills.

4) Scripts and awareness-heightening tools that help to sequence argumentation, scaffold

argument construction, and provide feedback that aids in modification of interactions between students.

5) Integration of multiple features that help guide students toward productive disciplinary engagement. For example, a specific database may be made available to students at an appropriate point in the process of argumentation that prompts its use.

The concepts just discussed are not only important to *learning* online, but also to the *assessment* of that learning, particularly in regard to formative assessment meant to be a part of the learning process. Effective instruction must be designed so that learning and assessment are connected (Black et al., 2004). Rowe and Asbell-Clarke (2008) conclude that effective online course design must be student-centered and that assessment of learning derived from online activities, such as discussions and scientific argumentation, may need to have different success criteria compared with traditional assessments. These criteria may diminish the importance of right and wrong answers while emphasizing the extent to which students accurately take on authentic scientific roles when creating and critiquing claims (Ford & Forman, 2006). For example, the assessment of what Ford and Forman (2006, p.3) call the “grasp of practice” or “…the familiarity with how the discipline decides upon knowledge claims,” if made measurable, may be a valid way to determine the level of student skill in scientific argumentation. In other words, if familiarity with the use of Newton’s Laws is evidence of skill in solving mechanics problems, grasp of practice may be used as evidence of skill in scientific argumentation. This concept can apply to both online and traditional instruction.

Methods used to evaluate argument quality, whether online or in traditional settings, vary from study to study. Importantly, research that skews toward explicit instructional strategies that emphasize argument structure, such as Zohar and Nemet (2002), do not necessarily limit evaluation strategies to structural analysis. Likewise, investigators who privilege immersion strategies, as do Bell and Lynn (2000), often use evaluation frameworks that rely heavily on the presence or non-presence of the normally accepted components of argument structure. Although there are some common characteristics that many frameworks evaluate, such as having a claim or assertion, or some manner of justification, the goals of the research most often guide the assessment structure and help define what counts as an important feature of a well-crafted argument (Sampson & Clark, 2008). Depending on what the framework is designed to measure, argument quality can vary from low to high for the exact same construction. Each approach to argument evaluation provides a different insight into how students argue and what proves difficult for them in the process (Sampson and Clark, 2008).

Common assessment frameworks use structural analysis based on a model of argument suggested by Toulmin (1958). These evaluative schemes generally look for a claim followed by evidence (data) and justification (rationale, based in accepted theory). Some analyses, like those used by Osborne, Erduran, and Simon (2004), look for other features of that model such as backings, qualifiers, and rebuttals. Such frameworks are domain-general in that they show little regard for accurate and appropriate content. Some high-quality arguments, as judged by domain-general frameworks, aren't valid when content is closely examined. Sampson and Clark (2008) suggest that there are domain-specific frameworks that can address the content concern, but are not without their own problems. Among these are analytic systems that either don't uniformly apply content importance to all aspects of the structure, or are effective but not generalizable due to excessive specificity. An example of the latter is found in the work of Sandoval and Millwood (2005), where evaluation of argument quality assays both content and structure. In their framework, content is judged through discipline-based conceptual quality and sufficiency of evidence, while structural quality is measured by what they term *rhetorical reference.* Essentially, rhetorical reference evaluates the skill with which students use data.

Clark et al., (2007) provide an overview of various frameworks that have been used in recent research to assess dialogic argumentation for online learners. They claim that challenges will always exist when measuring student success in online argumentative practice, but that the chosen framework should "…reflect specific perspectives on argumentation, pedagogical goals, and environmental structures" (Clark et al., 2007, p. 345). In other words, argumentation assessment frameworks are not "one-size-fits-all." For example, if the goal of learning is for a student to apply their knowledge of argument structure across various disciplinary areas, the assessment of argument quality is independent of the content of the argument. In fact, as mentioned above, when judging by structure alone, a high-quality argument may contain invalid content. Here is a short example of an argument two students may have when discussing a physics experiment on the topic of free-fall. The argument has good structurally quality because it contains a number of important argument components such as claims, observed evidence, and justification. However, the final claim is unjustified.

> Student 1: "I see that when I toss a rock upward, it stops at the top and returns to my hand so I can catch it." [*observational evidence*]
>
> Student 2: "That's because it underwent a constant acceleration due to gravity." [*valid claim followed by valid justification*]
>
> Student 1: "I agree that gravity pulls the rock back down, but the acceleration can't be constant because the direction of the motion changes." [*valid claim followed by invalid claim followed by observational evidence*]

Student 2: “Oh yeah…that’s right. Because the rock stops and turns around, the acceleration must change. [*observational evidence followed by invalid claim*]

The students co-construct this argument using acceptable components of argumentative practice, but conclude something that is false. This example shows the importance of the sequence of the contributions. It would be fine if the non-normative claims were followed by contributions that were normative, but in this case, the reverse was true. Thus, if a goal of the activity is to learn physics principles, the assessment of the argument must include more than the presence of structural components. Clark et al. (2007) describe additional frameworks that may be appropriate in a given circumstance based on the educational goal of the activity and the learning environment. Such things as conceptual quality, the nature and function of contributions, the epistemic nature of reasoning, and argumentation sequences and instructional patterns have all been used to judge the quality of online argumentation (Clark et al., 2007).

2.1.8 Assessment Tied to Learning Progressions

Pelligrino (2013) argues that assessments themselves should be evidence-based and that the results of an assessment should place a student on a continuum that identifies her level of proficiency. This approach recognizes learning as a progression toward greater understanding. Berland and McNeill (2010) suggest that carefully designed learning progressions are important when students are developing argumentation skills. They claim that learning progressions related to argumentative practice can be identified along three dimensions that call forth ever-greater complexity in student performance. These include the context within which the instruction occurs, the product that results, and the process through which it is created. It is important that assessment designers build in specific performance measures that provide evidence for claims of proficiency along such progressions (Pelligrino, 2013).

For example, Gotwals, Songer, and Bullard (2012) conducted a study with elementary school students designed to investigate an assessment system's ability to determine if students can learn how to blend course content with authentic science practice. The framework used rubrics that correlated to both content and practice learning progressions so that students could be evaluated as exhibiting some level of partial skill development (they used the term "middle knowledge") as well as mastery of both topic and task. Table 3 shows the general scoring rubric that was used to interpret a student argument. Student work was interpreted as being at one of four levels. A customized version of this rubric was used for each task that also included a level zero, indicating that the student made an inaccurate claim. Cognitive interviews supplemented the findings to assess the task process from the student perspective.

Table 3. General scoring rubric used by Gotwals, Songer, & Bullard (2012)

Level	Description
Level 4	Student constructs a complete evidence-based explanation (with an accurate claim, appropriate and sufficient evidence, and reasoning).
Level 3	Student makes an accurate claim and backs it up with appropriate and sufficient evidence but does not use reasoning to tie the two together
Level 2	Student makes an accurate claim but does not back it up with evidence or reasoning claim and backs it up with insufficient or partially inappropriate evidence
Level 1	Student makes an accurate claim but does not back it up with evidence or reasoning

Several important findings were reported. Gotwals, Songer, and Bullard (2012) claim that the interpretation of student performance is evaluative and can be difficult, especially for tasks such as argument creation as an open-ended task. An evaluation rubric is not always simple to use, as student-created arguments can show great variety, such as a claim with only weak evidence, a claim with only justification and no evidence, a claim without either evidence or justification, or a claim with evidence confounded with justification or visa-versa (when using a scaffold), among others. The authors indicate that, due to the diversity of student performance and expression of their understandings, rubrics such as theirs will not be able to interpret all student responses. The majority, however, they claim can be interpreted with validity. It is the messy nature of middle knowledge, exhibited when students know some aspects of a topic but have not mastered it, that makes this task difficult. Rapanta, Garcia-Mila, and Gilabert (2013, p. 488) make a similar point regarding assessment of argumentative practice in that "…both analytical and evaluative aspects of argumentative competence are considered problematic…Not only must they choose among aspects to focus on, but they also must ensure that the selected appraisal criteria are valid and reliable, meaning that they measure what they are supposed to measure in a repeatable and systematic way."

Sampson and Clark (2008) suggest future research is needed with evaluation frameworks that are holistic in approach rather than atomistic. These may provide greater insight into effective instructional strategies designed to support a student when learning argumentation skills. Regardless of the acknowledged difficulty of assessing student-constructed scientific arguments, the results gleaned by those used in educational research of the past few decades have provided a great deal of insight about this topic (Sandoval & Millwood, 2005), yet there is much more to learn (Kuhn, 2012).

2.2 Literature Review Summary

This review provided a foundation on which to design a research project whose findings led to insights regarding the problem of practice I identified as an online AP Physics 1 instructor.

Namely, online AP Physics 1 students must be proficient in scientific argumentation even though appropriate, practical, and effective instructional strategies and assessment frameworks have not been definitively established. Several important building blocks were provided by previous research on argumentation. These included the following:

1) Understanding the epistemic value of argumentation is important if students are to engage in its practice. The classroom culture should engender this understanding.

2) Instructional strategies that hybridize explicit instruction and immersion in practice may provide the best opportunity for students to appropriate argumentation skills.

3) Instructors should clarify the goals of instruction in, and assessment of, argumentation skill to create effective learning environments that foster their development in students.

4) Assessment of student performance vis-à-vis argumentation is a complex task best accomplished via well-thought-out rubrics that account for both structure and content.

3.0 Methodology

Here I outline the methods, instruments, and analyses used in this investigation. The development of the inquiry in general, and the data collection instruments in particular, were guided by the goals of the investigation (see section 3.3) and the information needed to address these inquiry questions:

Inquiry Question 1

What is the incoming level of argumentation skill of my students?

Inquiry Question 2

What patterns of progress in argumentation skill development do students exhibit as they engage in the constellation of argumentation activities offered at Physics Prep?

Inquiry Question 3

Are the resources offered at Physics Prep sufficiently robust to provide students the opportunity to reach levels of argumentation skill needed to effectively respond to assessment items they will encounter on the AP Physics 1 examination?

3.1 Inquiry Setting

The inquiry took place with students enrolled in an online AP Physics 1 course developed by me at a company I founded in 2012 called Physics Prep, LLC located in southwestern Pennsylvania. I offer four different Advanced Placement (AP) Physics courses to high school students. These include two algebra-based AP Physics courses called AP Physics 1 and AP Physics

2, and two calculus-based courses called AP Physics C Mechanics and AP Physics C Electricity and Magnetism. The Physics Prep website has provided resources and guidance for hundreds of students from around the world to either independently follow a defined workflow or to participate as homeschoolers through a partner organization, Pennsylvania Homeschoolers. The homeschool students attend bi-weekly live online instructional sessions and receive teacher feedback for submitted assignments as supplements to their independent work. Additionally, there is an active online course discussion forum on which the homeschool students are encouraged to ask and answer questions. Some, but not all, students also participate in peer study groups (via video conferences) designed to help with problem solving and conceptual development. They also earn a course grade on an official transcript that provides credit toward high school graduation and college application.

At Physics Prep, the AP Physics 1 course can be taken as a first semester accelerated course, allowing students to take AP Physics 2 in the second semester, or as a full-year course. There is no difference between the two options other than the pacing of the course. Since the design for each of the AP Physics courses was constructed by the College Board to match university level expectations, their content and learner outcomes are specific and extensive (College Board, 2012). For example, in addition to subject-related proficiencies, the College Board requires that students become skilled in constructing and responding to scientific claims based on justified evidence.

There are several instructional strategies used at Physics Prep that attempt to develop these skills in students. These include discussing and practicing argumentation skills during live sessions, requiring their use for inquiry-based lab activities, and assessing student responses to false scientific claims that I propose in every unit. Most importantly, I also require students to respond to open-ended unit test assessment items that require argumentation skills for successful completion. These free-response questions are similar to those found on the AP Physics 1 examination. For example, a student may be asked to make an evidence-based claim about a given scenario along with theoretical justification indicating why the evidence is appropriate for supporting the claim. Extensive feedback is provided by me for each of these instructional activities, however, peer-to-peer interaction is not as common.

The online inquiry setting provides the affordances of easy data collection from generally eager-to-learn students along with a platform wherein activities designed to develop specific skills can be carefully and consistently planned, enacted, and assessed. Electronic submission of assignments, along with associated feedback mechanisms, make communication between my students and me to be easy and trackable. The setting also includes common constraints associated with online education. These include a lesser degree of social interaction than might be optimal for the development of argumentation skills and a lack of day-to-day interactivity between the students themselves and between me and the students. The latter presents a challenge when particular students have not turned in assignments per the due dates provided on the pacing guides.

In general, the affordances and constraints affect learning differently for each student. For one who is self-motivated and has at least one highly-involved parent or guardian, I believe that the communication structures present in this online course are more than adequate to promote success. However, for students who lack motivation, become overscheduled, succumb to illness, or lack strong parental support, the chance of falling through the cracks is a real concern. For these and other reasons, it’s not uncommon for students to register for the course and then drop out after falling behind even though extensive efforts are made to accommodate for individual needs. These include one-on-one video chats, phone calls to parents, personal tutoring by a teaching assistant, peer-tutoring groups, assignment feedback, and alternative assignments and re-tests when appropriate.

3.2 Stakeholders

The three stakeholders associated with this inquiry are me, my students, and the parents of my students. As the owner, developer, and instructor at Physics Prep, I have a deep personal interest in making the courses I offer as effective as possible. This interest springs from several sources, the most important being my desire to increase access to high-quality AP Physics instruction to students who otherwise wouldn’t have it. The inquiry topic of argumentation is highly associated with this motivation. The College Board’s emphasis on the development of evidence-based argumentation skills for its algebra-based AP Physics courses is different from its previous stress on computation and broad coverage. Without purposeful adaptation of instructional strategies in this regard, I assume that course quality and effectiveness would suffer, making the goal of offering high-quality courses unlikely met.

The students who enroll in my courses range in age from thirteen to eighteen years old. The younger pupils are accelerated learners who have taken high levels of mathematics earlier than is typical for most students. While the majority of the students live in the United States, typically fifteen percent reside in other countries. Each student must complete an application process, including a math-readiness test that assesses algebra, geometry, and trigonometry skills. Parents (or guardians) guide the educational process of homeschoolers and take responsibility for meeting state graduation requirements instead of the local school district or a private school (Kunzman, 2012). Often, parents provide instruction in early grades, and find tutors or participate in family networks for instruction in disciplines where they lack expertise (Hanna, 2012). Thus, parents are important stakeholders in this inquiry in that their supportive task is made more difficult when high-quality resources are not available.

Hanna (2012) reports that the use of technology and online instruction by homeschoolers in Pennsylvania is growing. This is consistent with the enrollment statistics at Physics Prep which has shown growth in each school year since its inception. Over the past two years, approximately eighty students have enrolled in AP Physics 1 at Physics Prep through Pennsylvania Homeschoolers. Since the application process does not require socioeconomic information, no breakdown by socioeconomic status is available. However, in order to participate, students must have access to a computer and the internet.

Based on anecdotal evidence gleaned through personal conversations and other interactions, I can report that the majority of my homeschool students have taken several, if not many, online courses in their academic careers and are typically comfortable with the social aspects of the live online instructional sessions. They also frequently post questions the class discussion forum.

3.3 Inquiry Design

This inquiry was designed as *action research*. Action research is a sub-type of practitioner research wherein the investigator is engaged in “structured self-reflection” on her own practice (Edwards & Talbot, 1999, p. 61). Denscome (2014, p.120) describes action research as focused on “practical issues – the kinds of issues and problems, concerns and needs that arose as a routine activity ‘in the real world’.” *Teacher action research* is recognized as one of three methodologies that fall under the general category of action research, the other two being *practical action research* and *participatory action research* (Essays, 2018). In teacher action research, the classroom is used as the experimental setting with the goal of improved instruction that leads to improved learner outcomes. Although there are advantages to teacher action research such as the introduction of relatively fast positive change, a critique is that it can be overly subjective and plagued by personal bias.

Action research is characterized by four features. These include having (1) a practical nature, (2) an emphasis on changing the system being investigated, (3) an iterative character that allows for adaptation during the research, and (4) the active participation of practitioners (Denscome, 2014). My inquiry fits this description very well. I am interested in systematically investigating possibilities for change within my practice to provide benefit for my own students. My working hypothesis was that specific types of purposeful instruction designed to develop argumentation skill within a classroom culture that values argumentation will effectively prepare my students for free-response questions on the AP Exam that require their use. These specific instructional designs evolved throughout the investigation based upon ongoing data analysis. Importantly, I conducted live online sessions bi-weekly (twelve during this investigation) that set the learning context as one that values the construction of valid scientific arguments. This was done through several instructional strategies including, (1) the direct instruction of argument structure, (2) the engagement (immersion) of students in argumentation activities that demand deliberation, and (3) student claim-making and critique.

The choice for this design rests on three important factors. The first is that as artifacts were analyzed, the areas of instructional emphasis and learning context were adapted in response to the findings. This flexibility is a strength of action research. The second is that the nature of the inquiry setting, a single-person company, allowed for change to occur quickly and easily. This characteristic is in line with the practical aspect of action research. Third, as the practitioner whose workplace was under investigation, I had a strong interest in improvement as it relates to myself and my students.

I had two goals in mind when conducting the investigation, both of which are related to the prediction that my students will learn to construct strong scientific arguments. The first was that students would improve their skill in *disputative* argumentation whose purpose is persuasion. This is a skill that has broad benefits, extending beyond the classroom. Having skill in disputative argumentation allows a citizen to actively and effectively participate in public debate of socio-scientific topics. The second goal was to enhance student skill in *deliberative* argumentation. Deliberative argumentation privileges sense-making over persuasion. It is required for productive cooperative work in science practice. Interestingly, it’s also important when effectively responding to open-ended questions on science assessments, when students internalize the dialogic process, acting as both *constructor* and *critiquer* of claims in search of a sensible response produced in written form.

3.3.1 Overview

Data collection for this investigation took place in three phases: (1) the precourse phase, (2) the artifact collection phase, and (3) the poststudy phase. Collection of data concluded after unit five of the seven-unit AP Physics 1 full-year course, so the final phase could not correctly be called the “postcourse phase”. Table 4 presents how the investigation’s goals and inquiry questions are related to the data collection instruments used and the type of analysis performed on the data (qualitative, quantitative, or both). A rationale for each instrument is then discussed.

Table 4. Matrix of data sources, goals, and inquiry questions

Measured by…	Research question(s) addressed	Goal 1: Improve the disputative argument skill of my students	Goal 2: Enhance the deliberative argument skill of my students
Precourse Test	1	✓ quantitative	
Precourse Cognitive Appraisal Interview	1	✓ qualitative	
Student Artifacts from Lab Reports	2	✓ quantitative	
Student Artifacts from Addressing False Claims	2	✓ quantitative	
Student Artifacts from Online Discussion Forum	2	✓ qualitative	✓ qualitative
Student Artifacts from Free-Response Questions on Unit Tests	2,3	✓ quantitative	✓ quantitative
Poststudy Test	3	✓ quantitative	
Poststudy Cognitive Appraisal Interview	3	✓ qualitative	

Pre/poststudy comparisons are often used to develop findings in studies focused on argumentation. In one instance, Bell and Linn (2000) used this method to study how argument building correlated to knowledge integration. In another example, Iordanou and Constantinou (2015) compared preintervention dialogic argumentation artifacts with those constructed post-intervention to measure student progress. McNeill (2009) used the pre/posttest method to investigate the changes in student argumentation construction as related to teacher support. In this study, pre/posttest data will be used to measure change in student argumentation skill over time.

Analysis of student-created artifacts is also a common practice in research done on argumentation (Manz, 2015). Artifacts can take many forms including written arguments (Osborne, Erduran, & Simon, 2004) and transcripts of group discussions (Bell & Linn, 2000). Often, the argumentation activity through which the artifact is produced is supported by a learning scaffold such as a writing frame. In its simplest common form, analysis of argument quality for artifacts is done via a rubric using the primary aspects of Toulmin's basic argument model: claim, evidence, and justification (Toulmin, 1958). Additionally, artifacts can be evaluated for content accuracy (Gotwals, Songer, & Bullard, 2012). In this investigation, several types of student-constructed artifacts were evaluated for structural and content quality in order to determine if patterns of progress in argument skill development existed (see section 3.4).

3.3.2 Participants

There were twenty-five students initially registered for the AP Physics 1 course prior to the start of the academic school year at Physics Prep in August, 2018. Twenty-four students (and parents) agreed to participate in the study. As the study progressed, nine of these students dropped the course for various reasons at different times. Fifteen students participated in the entire study, each of whom took the pretest, submitted assignments in each of the first five units of the course (student constructed artifacts), and took the posttest. Additionally, six of the fifteen full-study students were interviewed both prior to the start of the course and after the study concluded. The data discussed in this report pertains to the fifteen full-study students only.

The participants were all homeschool students, each of whom had academic backgrounds that included algebra, geometry, trigonometry, and chemistry. Most of the students were from the United States, but there were three international students registered for the course. Each student had access to the internet and participated in the course by utilizing resources found at the Physics Prep website (Physics Prep, 2019), submitting assignments through Google forms, and accessing the course discussion forum. The students also participated in bi-weekly, live online sessions hosted through Zoom.com.

3.3.3 Conjecture Map

In order to develop a class culture that valued argumentation, instructional decisions were made to help promote the epistemic worth of arguing in a scientific manner. These included using direct instruction methods during recorded lectures and live sessions along with immersion methods during live sessions, discussion forum activities, and unit-based assignments. I conjectured that this hybridized approach would lead to learner outcomes that directly benefit students who take the AP Physics 1 exam. I believed that direct instruction on argumentation performed a normative function and was beneficial both cognitively and instrumentally. Additionally, I was confident that instruction through immersion activities, which took advantage of the sociocultural aspects of learning, was a powerful way to promote argument skill development. The authors of *A Framework for K-12 Science Education: Practices, Cross-Cutting Concepts and Core Ideas* (National Research Council, 2012, p. 73) state that, “Constructing and critiquing arguments are both a core process of science and one that supports science education, as research suggests that interaction with others is the most cognitively effective way of learning.”

So that the data collected during the investigation would have the most impact, analysis of artifacts produced in each unit were analyzed at six predetermined points in time (see section 3.4). This analysis allowed for both direct instruction and immersion activities to be adapted as the investigation unfolded. The unit by unit iteration of activities, and subsequent analysis, promoted pedagogical reflection that attempted to insightfully modify instruction in a strategic fashion.

The action research described in this study shares this iterative characteristic with another research methodology called *design research*. In a design experiment, evaluation of method is just as important as theory-making. In fact, the two evolve together in a repetitive fashion. Such an approach is warranted in educational research due to the complex and messy nature of instruction and learning (National Research Council (2002); Schoenfeld, 2006). Well-constructed action research and design research make use of the "Bayesian" assumption that new evidence can produce appropriate subjective adjustments to predicted outcomes. In other words, the more information, the better the adaptations. That new information is fed back into the system of instructional design so that theory and practice are intimately connected.

As opposed to traditional research methods, design experiments make use of what doesn't work as much as what does work (Gorard, Roberts, & Taylor, 2004). The same is true of action research. When analysis shows that some feature of a treatment is ineffective, this informs the adaptive work for the next iteration. Research of this type is free to evolve in the name of pragmatism and learns through its successes and failures. Unlike traditional research methodologies, interactions among participants and between participants and researchers can help guide the investigation.

Although there are clear benefits to research methods that include "on the fly" iterative analysis (Brown, 1992; Collins ,1992), both action research (Essays, 2018) and design research (Kelly & Lesh, 2002; Sloane & Gorard, 2003) have been criticized as lacking solid theoretical foundations. Although an overall proponent of design research, Dede (2004) is concerned that it is often "under-conceptualized" in that the conclusions drawn are often nothing more than common sense. In order to address this critique, as may be applied to this investigation, I have constructed a "conjecture map" that provides a foundation for articulating the theoretical basis of this study. Conjecture maps were proposed by Sandoval (2014) as a way to respond to critics of design research in general.

A conjecture map starts with a high-level conjecture that grounds the process. High-level conjectures are theoretically-based, general principles that help direct the design. For example, when designing the iPhone, Apple engineers were guided by a high-level conjecture claiming that if users were to find the device simple and easy to use, the design should reflect what people already know how to do (Kuang, 2011). Similarly, when designing an educational study, a research team cannot move forward in the process without an underlying principle they identify as important to learning. One such conjecture may be that "Scientific argumentation requires appropriation of discursive practices of making, justifying, and evaluating claims" (Sandoval, 2014, p. 25). The conjecture is not the design, but instead directs it, depending upon the context and focus of the research.

The design process continues as the high-level conjecture is *embodied* in context-specific aspects of a real-world situation under study, such as a classroom. Embodiments of a high-level conjecture include such things as explicit pedagogical structures, teaching materials, and planned discourse. Intra-study design conjectures are then articulated about observables (such as student artifacts) identified as mediating processes that emerge from the embodiments. These processes finally connect to expected learner outcomes through theoretical conjectures. The flow of the process is shown in the general conjecture map of Figure 1. As described, a conjecture map is a structure that outlines the commitment a research team has to investigating something worthwhile, and although it may evolve as the research proceeds, its construction provides the framework through which both theory and practice can improve.

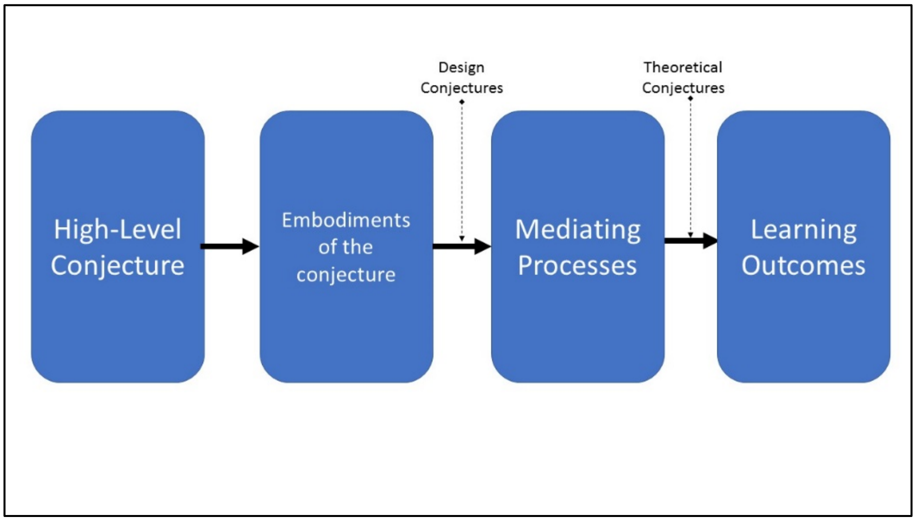


Figure 1. A general conjecture map (after Sandoval, 2014)

Conjecture maps provide links between aspects of the design, removing the sense that investigations are vague in their articulation. A conjecture map associated with this investigation is shown in Figure 2. This type of map allows one to distinguish between the conjectures made regarding research design, represented by the arrows in the center of the map, and those made regarding expected learning, represented by the arrows on the right side of the map (Sandoval, 2014). For example, the design conjecture, "If students make use of the course tools and resources, work on argument-centered assignments will result", is different from the theoretical conjecture, "If argument-centered assignments are completed, students will increase their appreciation of the epistemic value of argumentation." The conjecture map created for this study contains one high-level conjecture embodied by three specific aspects of the AP Physics 1 course. Those embodiments are followed by ten design conjectures from which three mediating processes emerge. Finally, six theoretical conjectures connect the mediating processes to learner outcomes. Even though there are two expected learner outcomes, only the first, *construction of strong* *scientific arguments*, was tested in this study. The second outcome, *increased appreciation of the epistemic value of argumentation*, is proposed as a valuable follow-up to this investigation.

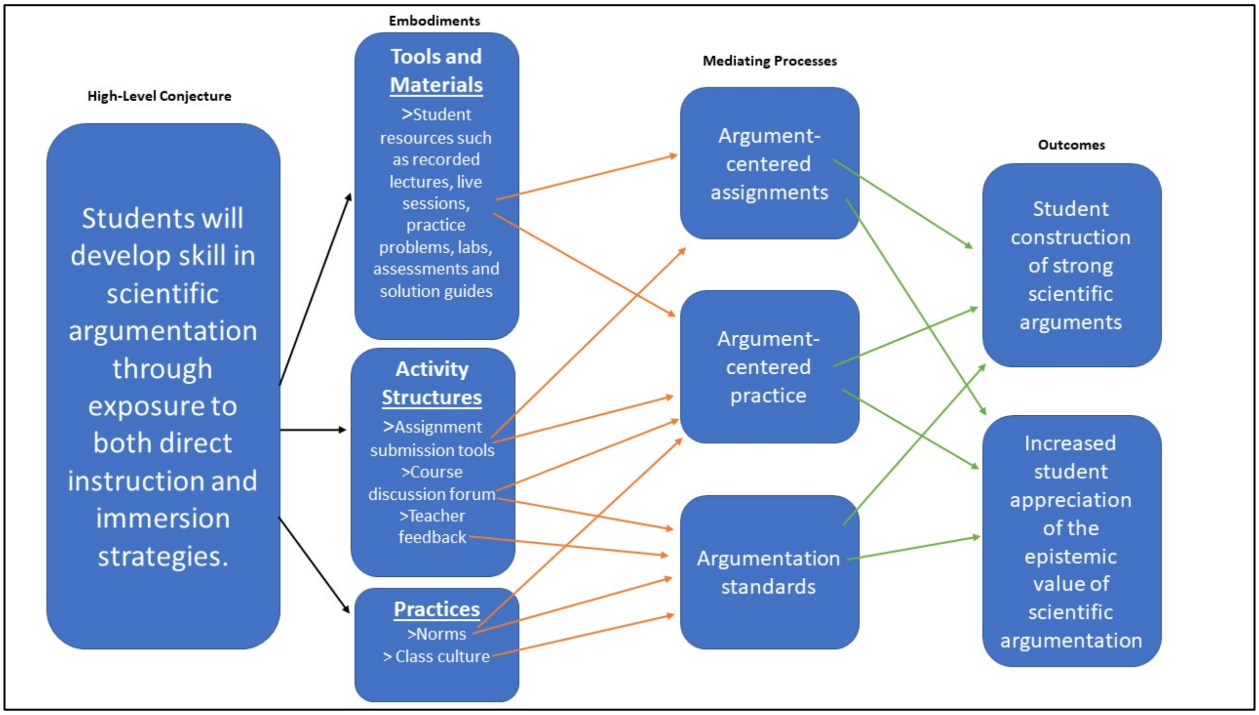

High-Level Conjecture
Students will develop skill in scientific argumentation through exposure to both direct instruction and immersion strategies.
Embodiments
Tools and Materials
>Student resources such as recorded lectures, live sessions, practice problems, labs, assessments and solution guides
Activity Structures
>Assignment submission tools
>Course discussion forum
>Teacher feedback
Practices
>Norms
> Class culture
Mediating Processes
Argument-centered assignments
Argument-centered practice
Argumentation standards
Outcomes
Student construction of strong scientific arguments
Increased student appreciation of the epistemic value of scientific argumentation

Figure 2. Inquiry conjecture map

3.4 Inquiry Methods and Evidence

The inquiry took place over a six-month time frame from late August, 2018 until February, 2019 - the conclusion of unit five in the seven-unit AP Physics 1 course. Six analysis points were identified wherein data collected to that point served to inform instructional decisions for live sessions as well as discussion forum activities. Table 5 presents a timeline of the analysis points.

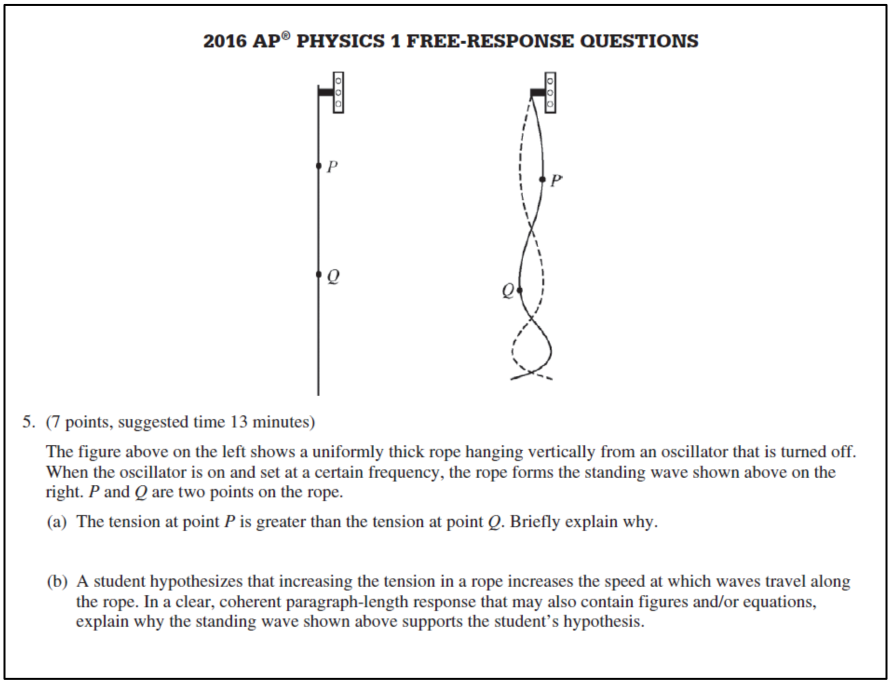
Table 5. Investigation timeline

Precourse	In-Course (Iterative)					Poststudy
• Test • Interviews	• Lab Report Argument Artifacts • False Claim Rebuttal Artifacts • Free-Response Argument Artifacts					• Test • Interviews
	Analysis Point 1	Analysis Point 2	Analysis Point 3	Analysis Point 4	Analysis Point 5	Analysis Point 6
Late August, 2018	Early October, 2018	Late October, 2018	Mid-November, 2018	Late December, 2018	Mid-February, 2019	Late February, 2019

At analysis point one, data from the pretest, the precourse interviews, and artifacts collected in unit 1 were examined to see if any general argument skill deficiencies or strengths could be identified. Based on that analysis, part of the instruction planned in the following live session was adapted to address the findings. Additionally, students were asked to participate in a discussion forum activity intended to clarify and strengthen a particular identified group weakness. Examination of student artifacts from units two through five was done at each successive analysis point in a similar manner, with instruction during the upcoming live sessions and discussion forum activities adapted to address the findings. Details of the instructional decisions are outlined in section 3.4.7.

3.4.1 Background

Prior to describing the proposed data collection instruments, a short description and example of an argument-related AP Physics 1 free-response exam question, published by the College Board, is discussed. The instruments used in this investigation can be compared to it. The first administration of the AP Physics 1 exam occurred in May, 2015. Each year the exam is composed of two equally weighted parts, a fifty-question multiple-choice section and a five-question free-response (FR) section. The questions in the FR section contain multiple parts (sub-questions). Every AP Physics 1 exam has one FR question that includes a “paragraph-length response” sub-question, such as the one shown in Figure 3 taken from the 2016 administration of the exam (College Board, 2016b), which requires the written construction of a scientific argument. This example requires a short-answer response in part (a) and a paragraph-length response in part (b). For paragraph-length responses, the College Board (2015b, p.1) requires “… a coherent argument that uses the information presented in the question and proceeds in a logical, expository fashion to arrive at a conclusion.” The other four FR questions often include short-answer type sub-questions that also demand skill in argument construction by asking students to explain, justify, or provide evidence to support their response.



2016 AP® PHYSICS 1 FREE-RESPONSE QUESTIONS

5. (7 points, suggested time 13 minutes)

The figure above on the left shows a uniformly thick rope hanging vertically from an oscillator that is turned off. When the oscillator is on and set at a certain frequency, the rope forms the standing wave shown above on the right. *P* and *Q* are two points on the rope.

(a) The tension at point *P* is greater than the tension at point *Q*. Briefly explain why.

(b) A student hypothesizes that increasing the tension in a rope increases the speed at which waves travel along the rope. In a clear, coherent paragraph-length response that may also contain figures and/or equations, explain why the standing wave shown above supports the student’s hypothesis.

Figure 3. FR question 5 - 2016 AP Physics 1 exam (College Board, 2016b)

In 2016, the paragraph-length sub-question was worth five points in the forty-five-point FR section of the exam. The argument-related short answer questions in that section were worth an additional twenty-six points. The importance of argumentation on the 2016 exam is clear, as 51% of the possible points in the FR section score were associated with that skill (see Table 6). For other exam administrations the argument-related percentage varied between 31% and 69% (College Board, 2019).

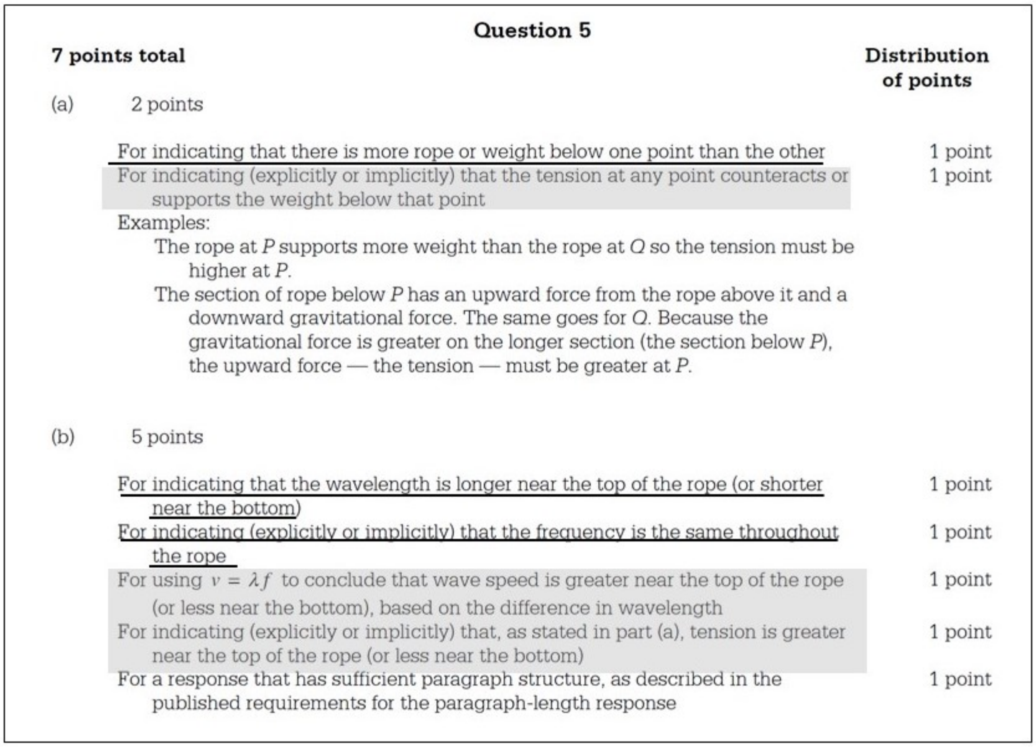
Table 6. Summary of argument-related items on AP Physics 1 exams (College Board, 2019)

Year	Total Number of Sub-Questions	Number of Argument-Related Short-Answer Sub-Questions	Number of Paragraph-Length Response Sub-Questions	Total Number of points	Number of points related to argument skills	Percent related to argument skills
2015	18	12	1	45	31	69%
2016	19	12	1	45	23	51%
2017	20	10	1	45	27	60%
2018	21	7	1	45	14	31%

The official scoring rubric for question five of the 2016 exam is shown in Figure 4. As is common for such scoring guides, the point-based structure for argument-related questions generally builds on the basic Toulmin (1958) notion that an argument is constructed of a claim, evidence, and rationale. In both parts of this question, a claim is given and the student must support it. On other AP Physics 1 exam administrations, an entire argument, including a claim (or a counterclaim) must be constructed. In either case, it’s clear that students benefit from being skilled at using evidence and rationale to support a claim. However, this type of rubric is not just looking for the construction of any argument. It is checking for the right argument.

In the scoring rubric, I have underlined evidence statements and highlighted justification statements (in gray). Statements of evidence point out something that can be observed or measured.

Justification statements are conceptual rationales that indicate why the stated evidence supports the claim. In this example, the justifications in both parts of question five use concepts associated with Newton's Second Law of Motion to conclude that the tension must act to counteract the downward-acting weight force. Additionally, the wave speed equation, a theoretical construction that connects wavelength and frequency to the speed of a wave, is used with the evidence provided to determine that wave speed is higher at the top of the rope. Each of the statements are worth one point. The final point in the rubric is awarded if the student structures the argument to conform with the College Board expectations of a coherent, logical exposition that arrives at a conclusion. In a similar manner, each of the data collection instruments described in the following sections of this report have evaluation rubrics that check for the structural quality of a student-produced argument and, when appropriate, its accuracy.



Question 5

7 points total	**Distribution of points**
(a) 2 points	
For indicating that there is more rope or weight below one point than the other	1 point
For indicating (explicitly or implicitly) that the tension at any point counteracts or supports the weight below that point	1 point
Examples: The rope at *P* supports more weight than the rope at *Q* so the tension must be higher at *P*. The section of rope below *P* has an upward force from the rope above it and a downward gravitational force. The same goes for *Q*. Because the gravitational force is greater on the longer section (the section below *P*), the upward force — the tension — must be greater at *P*.	
(b) 5 points	
For indicating that the wavelength is longer near the top of the rope (or shorter near the bottom)	1 point
For indicating (explicitly or implicitly) that the frequency is the same throughout the rope	1 point
For using $v = \lambda f$ to conclude that wave speed is greater near the top of the rope (or less near the bottom), based on the difference in wavelength	1 point
For indicating (explicitly or implicitly) that, as stated in part (a), tension is greater near the top of the rope (or less near the bottom)	1 point
For a response that has sufficient paragraph structure, as described in the published requirements for the paragraph-length response	1 point

Figure 4. Scoring rubric for FR question 5 - 2016 AP Physics 1 exam (College Board, 2016c)

3.4.2 Pre/Posttests

The pre/posttests given to the participants are discussed here. Both tests were composed of questions designed to determine the argument skill level of the study participants. Each was composed of the three questions and are shown in Appendix A. Commentary that explains why each question type was chosen along with example responses are provided. The purpose of the pretest was to help answer inquiry question one and to also provide a baseline from which to make a comparison to poststudy skill level.

In addition to these important investigatory purposes, knowing the incoming argument skill level of students was simply pedagogically valuable. Lawson (2004) claims that science teachers need to know the intellectual development level of students at the start of any course that requires scientific reasoning. He suggests that without sufficient intellectual development, students cannot construct scientific arguments or express scientific explanations. Efforts to construct arguments will "fall apart" unless students have developed their hypothetico-deductive reasoning abilities, which are present (but undeveloped) at birth, to sufficient levels (Lawson, 2004, p. 322). According to Lawson (2004), this occurs through an awareness that grows with time through reflection and application.

Using the stages of intellectual development described by Piaget, Lawson (2004) expresses that humans develop intellectually from the preoperational to concrete operational level of intellectual development sometime between the age of seven and preadolescence. At the concrete level, students can test hypotheses descriptively, but not causally. Causal hypothesis testing is associated with the formal operational stage of development which manifests in early to late adolescence. In order for a student to argue scientifically at the level required in AP Physics 1, she must have developed to the formal operational stage. Although Lawson couches his ideas in

Piaget's model of development, he carefully states that, "… use of the Piagetian stage labels does not imply acceptance of his theory concerning their underlying operations…" (Lawson, 2004, p. 323). Lawson's (2004) work represented a good foundation on which to base a pretest for my research. The pretest was meant to identify if students can construct scientific arguments. Therefore, items on that instrument consisted of various argument construction tasks that indicate human intellectual development up to and including the formal operational level.

Each of the items asked on the pre/posttests required a student to make a claim on given information and then fully explain why they made that claim. The student's response was used to determine whether she skillfully supported the chosen claim with evidence and justification (rationale) according to the rubric shown in Table 7. This rubric was modeled after that used by Gotwals, Songer, and Bullard (2012) but with the claim descriptor, *accurate*, taken out. The removal was made so that the score assigned to the response did not depend upon which claim the student chose. Instead, the score was based upon the level of support offered for any chosen claim. Cavagnetto and Hand (2012) make clear that the rationale for an argument does not have be made via completely separate declarations, but may be meshed within claim and evidence statements. However, in the rubrics developed by Gotwals, Songer, and Bullard (2012) that were customized versions of the general rubric designed to evaluate a specific task, responses that included implicit rationale were not evaluated at the highest possible level. For their customized rubrics, explicit reasoning was required to receive a level four evaluation. For purposes of this investigation, that requirement seemed appropriate based on expectations for sufficient argument structure as demanded by the College Board on AP Physics 1 exams (see Figure 4, p.61). Thus, I added the italicized word *explicit* shown in the level 3 and level 4 rubric descriptions for clarity and to emphasize the importance of constructing complete arguments that could be objectively analyzed to as high a degree as possible.

Table 7. Structure rubric

Argument Structure Evaluation	Description
Level 4	Student constructs a complete evidence-based explanation (with a claim, appropriate and sufficient evidence, and *explicit* reasoning that ties the two together).
Level 3	Student makes a claim and backs it up with sufficient and appropriate evidence but does not use *explicit* reasoning that ties the two together.
Level 2	Student makes a claim and provides evidence to support the claim. However, the evidence is either insufficient or inappropriate.
Level 1	Student makes a claim but does not back it up with evidence (even though a rationale may be attempted).

I expected that the use of the structure rubric with the pre/post test data would allow me to produce a snapshot in time at two important points in the study. They turned out to be just that; momentary views of the path an individual student was traveling amidst a jungle of competing concerns. With that in mind, I did not approach these assessments as ways to prove how lack of reasonable evidence or justification in the response was necessarily suggestive that the student ***could not*** support claims in general. Instead, I viewed them as indicative of how a student chose to respond to a particular question at a specific moment (Sandoval & Millwood, 2005). That said, if a pattern of not using appropriate evidence and explicit rationale throughout the pretest or posttest was noted, this may, among other possibilities, suggest that (1) the student was unaware of the requirements for creating a valid scientific argument, (2) was unable to analyze the given data for appropriate evidence, (3) assumed that the "audience" for which the response was intended shares a common understanding of warrant implicit in the response (Manz, 2015), or (4) was simply not interested in performing at the level she was capable of. In the first three cases, my working hypothesis (see section 3.3) was tested via a pre/posttest comparison. In case three, for example, improving argument skill may have been as simple as impressing upon students the importance of explicitly stating their reasoning in a way that logically connects the evidence to the claim and doesn't require assumptions to be made. In the fourth case, responses to pretest and posttest questions would lose meaning.

3.4.3 Pre/Post Cognitive Appraisal Interviews

The purpose of the cognitive appraisal interviews was to have the interviewee reveal the process through which they respond to prompts that require skill in argumentation. As with the examination of pretest data, the analysis of precourse interview data helped to determine the in-coming argument skill level of the students. However, not all participants were interviewed. Due to time constraints following the pretest and the start of the course (one week), a purposefully selected sample of ten students were interviewed before the course began. Four of these interviewees did not complete the course through unit five, leaving six students for the poststudy interview. The initial ten interviewees were chosen based on pretest results by dividing the participants into three performance groups. Three precourse interview students were randomly selected from the lowest scoring group, four from the middle scoring group, and three from the highest scoring group. The results provided a baseline from which to make a comparison to poststudy skill level for the six remaining interviewees, one from the lower group, four from the middle group and one from the higher group. Analysis of the precourse interviews also helped to clarify instruction after analysis point one.

The protocols for the interviews are outlined in Appendix B. During the initial phase of the interview, students were asked to agree or disagree with four claims presented one at a time. The claims were not free of science content, but were grounded in concepts (density and thermal conductivity) that were likely, but not definitely, covered in previous prerequisite science courses. The students were asked to make use of data (shown to them in chart form) on several properties of different objects, along with related concepts listed next to the data, in their oral responses. They were also asked to continuously verbalize their thinking as they formulated their answers, which were recorded and transcribed for analysis. In the second phase of the interview, an additional row of data was added to the chart for a new object, with two of its properties missing. A question was then asked that required a response in the form of a written argument. The student was given as much time as needed to construct a response that was then read aloud. In the final phase of the interview, the student was asked to verbalize their thoughts on what makes an argument strong.

Cognitive interviews are a standard method used by researchers to allow "private speech" associated with survey-taking to be made public. Although used for different reasons by cognitive psychologists, cognitive interviews designed for research are most commonly used to test the validity of items on questionnaires and assessments (Silverman, 2010). Also called cognitive pretesting, Karabenick (2007) describes the use of this data collection tool as a way to determine if the cognitive processes designed into an item, and assumed to be used by the participant, matches that which is actually used. In other words, is the item validly measuring what it is designed to measure from a cognitive perspective? Silverman (2010, p.9) suggests that in addition to this important function, one type of cognitive interview, the cognitive appraisal interview, can be used, "…for gaining insight into participants' cognitive processes and interpreting alternative forms of data gathered as part of a broader, mixed-methods approach to research." So, the cognitive appraisal interview can be more than a tool to analyze other data collection tools. It can be used as a data collection tool itself. For my research, the cognitive appraisal interview added to the insights gleaned from the pretest through typological qualitative analysis based on that described by Hatch (2002) as suggested by Sampson and Blanchard (2012). Details of this analysis are outlined in section 3.5.3 of this report. Additionally, quantitative analysis was performed on each student interview response using the structure evaluation rubric (Table 7, p. 64). These evaluations helped to answer inquiry questions one and two.

3.4.4 Student-Constructed Artifacts: Lab Reports

The purpose of artifact collection was to document the argument skill level of individual students as they were exposed to purposefully designed teaching strategies. These strategies, and related activities, helped to create a learning context that valued argumentation. Analysis of artifacts helped to answer inquiry questions two and three. Scientific arguments constructed in lab reports are one type of artifact evaluated in this study and will be discussed here. Two other types, false claim rebuttals and argument construction on unit tests, will be discussed in later sections of this report.

In each unit of the investigation, lab reports were written by participants that demanded the use of argumentation skill. For example, in unit one, students measured the time it takes for a steel sphere to roll down a ramp at various angles. They also calculated the expected time for each trial through the use of kinematic equations. The time measurement was made on a stopwatch. The angle measurement was made with a protractor. As part of a lab report, students were required to choose from one of the two following claims and construct an argument to defend it.

Claim 1: The uncertainty in the measurement of time is larger than the uncertainty in the measurement of angle.

Claim 2: The uncertainty in the measurement of angle is larger than the uncertainty in the measurement of time.

The lab report artifacts were evaluated in two ways. The first was through assessment of argument structure, as was done for the pretest data, using the rubric shown in Table 7 on page 64. Additionally, the accuracy of the claim, and it's supports, were evaluated using the rubric shown in Table 8. Importantly, if the claim was not accurate, but accurate evidence and rationale statements were included in the argument, the accuracy evaluation was still zero. Sample responses are given in Table 9 showing both structure and accuracy evaluations. Again, claim statements are bolded, evidence statements are italicized, and rationale (justification) statements are underlined. Lab report argument assignments for all five units are listed in Appendix C along with samples of student responses.

Table 8. Evaluation rubric that checks for argument accuracy

Accuracy Code	Description
2	The claim is accurate and is supported by evidence and justification that is also fully accurate.
1	The claim is accurate but the evidence or justification is only partially accurate (or is missing).
0	The claim is inaccurate

Table 9. Example responses evaluated for structure and accuracy

Argument Structure	Sample Response	Argument Accuracy
Level 4	**The time measurement is more uncertain** *because the relative standard deviation of the time measurements (σ = 0.073 s, RSD = 4.7%) is greater than the relative standard deviation for the angle measurements (σ = 0.063 degrees, RSD = 1.2%).* <u>Relative standard deviation is an accepted measure of uncertainty</u>.	2
Level 3	**The angle measurement is more uncertain** *because the spread of the data is larger for the angle measurement (0.20 degrees) than it is for the time measurement (0.18 s).*	0
Level 2	**The time measurement is more uncertain** *because it's hard to stop the timer quickly but easy to use the protractor. This makes sense because humans have reaction times that vary from one person to another.*	2
Level 1	**The angle measurement is more uncertain than the time measurement.**	0

3.4.5 Student-Constructed Artifacts: Written False Claim Rebuttals

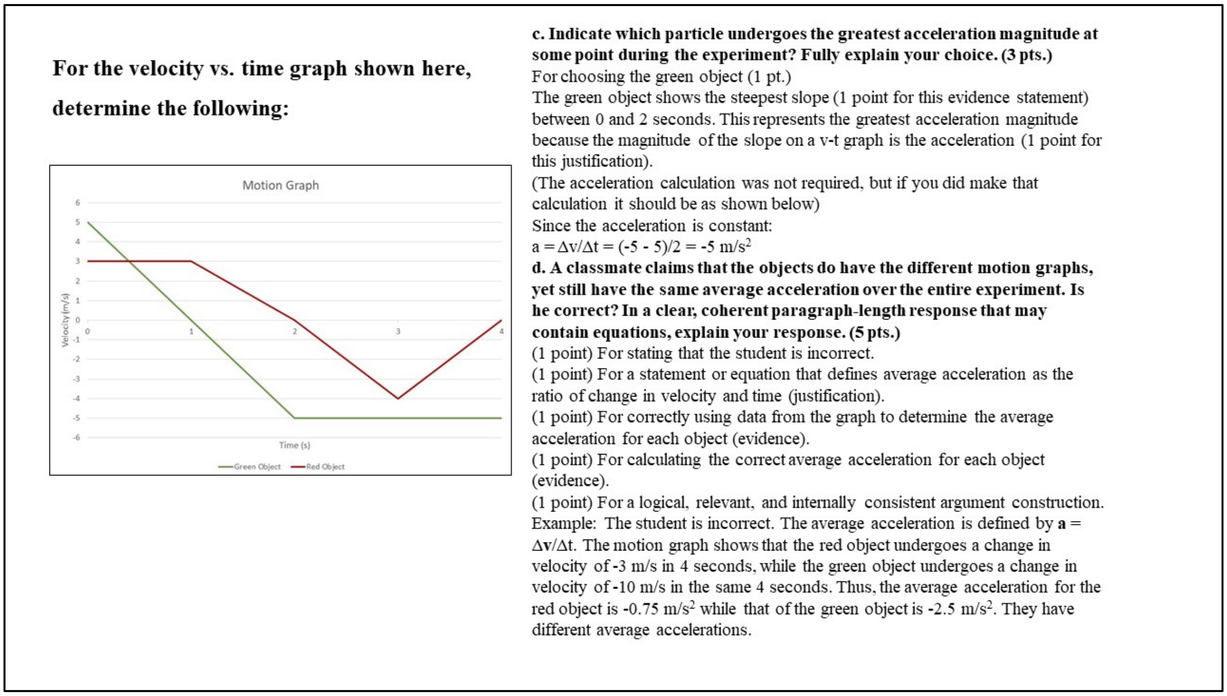
As part of review activities on the day prior to the unit test, students submitted rebuttals to a false scientific claim associated with the topics covered in that unit. For example, in unit one, students were asked to rebut the following false claim in writing: "When an object is tossed vertically upward, its acceleration vector cannot be constant because the object returns to the Earth." The other false claims used in the investigation, along with examples of student responses, are listed in Appendix C. The rubrics used to evaluate responses to false claims are the same as those used to evaluate lab report arguments. Table 10 shows sample responses with various combinations of structure and accuracy scores using the same component distinctions as above.

Table 10. Sample student responses for false claim example

Argument Structure	Sample Response	Argument Accuracy
Level 4	**The acceleration is constant.** *The measurement of velocity for the object will show that it changes uniformly each second, becoming more negative (by 9.8 m/s) each second. It slows as it rises, stops, then speeds up downward.* <u>This is because the force of gravity on the object is constant – and constant force produces a constant acceleration (F = ma and a = Δv/Δt).</u>	2
Level 3	**The acceleration is constant.** *The object stops at the top, and although the speed changes more quickly on the way up compared to on the way down, the object still undergoes the same overall change in velocity in both directions.*	1
Level 3	**The acceleration varies.** *You can measure the change of speed during the motion. The speed changes more quickly on the way up compared to on the way down.*	0
Level 2	**The acceleration is actually constant.** *You can watch the object to see how the speed keeps changing in both directions.*	2
Level 1	**The acceleration direction actually does vary for all objects thrown upward.**	0

3.4.6 Student-Constructed Artifacts: Arguments on Unit Exams.

On each unit test, one of the FR questions included the demand to construct a paragraph-length argument based upon information given in the question. The example shown in Figure 5, which contained four sub-questions (a) to (d), is from the test in unit one. Sub-questions (c) and (d) both required the student to use argumentation skills, but only part (d) required a fully constructed scientific argument. The scoring guides used to assign point values to the students' answers are shown beneath argument-related sub-questions in Figure 5. This type of rubric is modelled after those used by the College Board to evaluate FR questions on AP Physics 1 examinations and was used to help answer inquiry question three. Additionally, like the lab report arguments and the false claim rebuttals, responses to the paragraph-length questions were also analyzed using the structure and accuracy rubrics in Tables 7 (p. 64) and 8 (p. 68). This allowed three student-constructed artifacts (a lab report argument, a false claim rebuttal, and a paragraph-length FR argument) to be evaluated via those rubrics, so that an average structure and accuracy score could be assigned to each participant in each unit. The FR questions used in all five units, along with their scoring guides, are shown in Appendix C.



For the velocity vs. time graph shown here, determine the following:

Motion Graph
Velocity (m/s)
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
0
1
2
3
4
Time (s)
Green Object
Red Object

c. Indicate which particle undergoes the greatest acceleration magnitude at some point during the experiment? Fully explain your choice. (3 pts.)

For choosing the green object (1 pt.)

The green object shows the steepest slope (1 point for this evidence statement) between 0 and 2 seconds. This represents the greatest acceleration magnitude because the magnitude of the slope on a v-t graph is the acceleration (1 point for this justification).

(The acceleration calculation was not required, but if you did make that calculation it should be as shown below)

Since the acceleration is constant:

$a = \Delta v/\Delta t = (-5 - 5)/2 = -5\ \mathrm{m/s^2}$

d. A classmate claims that the objects do have the different motion graphs, yet still have the same average acceleration over the entire experiment. Is he correct? In a clear, coherent paragraph-length response that may contain equations, explain your response. (5 pts.)

(1 point) For stating that the student is incorrect.

(1 point) For a statement or equation that defines average acceleration as the ratio of change in velocity and time (justification).

(1 point) For correctly using data from the graph to determine the average acceleration for each object (evidence).

(1 point) For calculating the correct average acceleration for each object (evidence).

(1 point) For a logical, relevant, and internally consistent argument construction.

Example: The student is incorrect. The average acceleration is defined by $\mathbf{a} = \Delta\mathbf{v}/\Delta t$. The motion graph shows that the red object undergoes a change in velocity of -3 m/s in 4 seconds, while the green object undergoes a change in velocity of -10 m/s in the same 4 seconds. Thus, the average acceleration for the red object is -0.75 m/s^2 while that of the green object is -2.5 m/s^2. They have different average accelerations.

Figure 5. FR question from unit 1 test

As was seen in Tables 9 (p. 68) and 10 (p. 69) for lab arguments and FC rebuttals, FR submissions could have combinations of structure and accuracy scores that don’t necessarily correlate with each other. In other words, FR arguments with a high structure scores don’t have to have a high accuracy score. For example, if a student created an argument based on a force analysis, but failed to identify all the forces acting on an object, a claim about acceleration would be inaccurate even though the evidence cited was appropriately used in Newton’s second law. In that case, an inaccurate claim was structurally well-supported.

3.4.7 Hybridized Instruction on Argumentation

A combination of direct instruction on argumentation and instruction through immersion was planned and implemented in the course during the investigation. Top-down, teacher to student didactic discourse was used to introduce and reinforce the learning of argument structure via two important course activities: recorded lectures and live instructional sessions. These generally focused on baseline knowledge needed to construct a scientific argument and how it differed from common everyday argumentation. Additionally, activities that prompted student involvement in the creation and critique of scientific arguments were designed for each unit in the course. These immersion activities took place during live sessions, through student posts (and responses to posts) to the class discussion forum, and through assignments that required the construction of an argument. Appendix D presents details about when and why these strategies were used. With the exception of the first three occurrences, instead of following a rigid instructional plan, live session activities associated with argumentation practice were adapted to respond to deficiencies noted in work submitted by students from the weeks prior to those sessions. This meant that analysis and discussion of selected student-constructed artifacts were inserted into the instructional activities when needed. As the course proceeded, and student skill evolved, emphasis was made on various areas of argument construction, but always stressed the importance of including basic structural components to produce a high-quality scientific argument. Immersion in argumentation in this online course was not as robust as might be accomplished in a traditional classroom where opportunities exist daily for student-student and student-teacher interaction. However, care was taken so that when occasions arose that allowed for such interaction, it was fostered. The combination of direct instruction of argument structure and the immersion of students in argument creation and critique was meant to promote a class culture in which the value of scientific argumentation was clear and continuous.

3.5 Analysis and Interpretation

This section will discuss how data was analyzed and how that analysis helped to inform the instructional content at various points in the investigation. I will provide an overview of both the quantitative analysis used to evaluate every student construction collected in the study and the qualitative analysis performed on the interview data. Finally, I will also describe the role of second coders for each data collection instrument.

3.5.1 Continuous Use of a Research Journal

As the investigation unfolded over a six-month period, careful recording of the details of the study, along with the data used at each analysis point, were kept in an electronic journal. When trends in student work were identified through journaling, they were used to inform upcoming instructional plans as described in section 3.4.7. Insights were recorded that helped me to clearly recall the issues that were apparent at the time. Specific student artifacts were referenced that acted as exemplars of the progress, or lack thereof, exhibited by my students in regard to argument skill development.

Additionally, the research journal supplied a convenient space for overviews of pertinent literature not included in the initial literature review. These laid a foundation for the new insights noted at the analysis points. In this way, a sense of confidence developed as the study progressed.

Work by Schoenfeld and Herrmann (1982) and Chi, Feltovich, and Glaser (1981), was particularly important, as it provided backing for my emerging intuition about the role that conceptual understanding plays in problem representation and for my purposes, argument construction. This link will be explored in chapter five. The journal allowed the practical and the theoretical to mix together on its pages. Some entries in the journal led me down dead-ends while others spawned ideas that found their way into this report. At the end of each section of the journal, a list of actions steps was written that gave structure to the investigation in the weeks thereafter. I thought of these as the aspects of the study design that allowed the investigation to be appropriately called *action research*.

3.5.2 Coding of Quantitative Data

Each data collection instrument used in this study produced items that could be evaluated either by using the structure rubric (Table 7, p. 64) alone or in combination with the accuracy rubric (Table 8, p. 68). The structural evaluation was done in three parts. First, the item was color coded to identify the argument components it contained. Those components were then judged against the structure rubric to obtain a score of one to four. Finally, for evaluations that were not at level four, comments were written to explain what was missing in the argument. Often, these comments were forwarded to the student constructor for formative assessment purposes. For the items that were also evaluated for accuracy, each of the components were judged to be scientifically accurate or not, and a score from zero to two was determined using the accuracy rubric. Examples of the process are presented in Table 11 for two arguments made in response to the following teacher-created false claim: “When an object is tossed vertically upward, its acceleration vector cannot be constant because the object returns to the Earth." Claims are in bold font, evidence is italicized, and justification is underlined.

Table 11. Examples of structure and accuracy evaluations

Argument	Comment	Structure Score	Accuracy Score
(Student 948) **This claim is false**. The reason why this claim is false lies within the question itself. *The question says the object returns to the earth*. The reason the object returns to the earth is gravity. Gravity acts upon every object with the same amount of force -9.8m/s^2. This is the acceleration of the object. This acceleration is constant. Therefore, *every object tossed vertically up while on earth has a constant negative acceleration (-9.8m/s^2).*	You don't discuss much evidence (anything you can see or measure) in your argument except that the object returns to Earth. You do give the acceleration at the end, but not enough description of what that would involve. What else would you see that would indicate that the object has a constant negative acceleration? Also, the acceleration due to gravity (-9.8 m/s^2) is not a force as is stated in your argument.	2	1
(Student 886) **The tossed object's acceleration is constant.** *Its initial velocity would have a high speed then will start slowing down as it reaches its maximum height. At its maximum height, the object will temporarily stop and then speed up as it falls back down to the ground.* Though there is change in the object's velocity, *its acceleration is constant at -9.8 m/s^2.* This is because an object in free fall conditions will have a constant acceleration because of gravity, regardless of the initial velocity.		4	2

Analysis of the quantitative data was done in two ways. In addition to graphical and chart-based analyses that visually showed data at specific points in the investigation or presented changes over time at both the individual and group levels, statistical analyses (paired-sample T-tests) were performed for all appropriate data to help determine effect size using Cohen's d. The standard effect size levels (low > 0.2, medium > 0.5, high (>0.8) were used to interpret the findings. Due to the small size of the sample, statistical significance, while reported, was not discussed in the analysis to avoid making claims that may not be true. Pearson correlations were used to find possible associations between pretest and precourse interview results, structure and accuracy evaluations, and FR structure and point-based evaluations. The strength of association was judged as weak for $0.1 < |r| < 0.3$, moderate for $0.4 < |r| < 0.5$, and strong for $0.5 < |r|$.

The results of the analyses were used to help answer all three inquiry questions (see section 3.0). For inquiry question one, the incoming argument skill level of students (as a group and individually) was judged based on precourse data relative to the standards presented in the structure rubric (Table 7, p. 64). For question two, patterns of progress were examined through graphical and chart-based means while changes in student argument skill level were evaluated by statistical tests. Question three, regarding the sufficiency of course resources relative to published AP Physics 1 argument-related expectations, was answered by comparing precourse and poststudy assessment results and by comparing average student performance on unit test items that required argument construction against evaluation standards set by the College Board. The College Board reports AP test scores via five categories, the highest of which indicates "extremely well-qualified" status. Because the AP Physics 1 exam is graded on a significant curve, a score of more than seventy-percent on the exam generally associates with that status. This cut-off percent changes from year to year, but has remained close to that level since the inception of the AP Physics 1 course in 2014. The number of students who earn that highest status each year is relatively small, with only about 5% of AP Physics 1 test-takers placed in that category.

3.5.3 Qualitative Analysis of Interview Data

A typological analysis of the responses to precourse and poststudy interview questions was completed based on Hatch (2002) using NVivo qualitative analysis software. The interview protocols can be found in Appendix B. The analysis began by identifying the typologies to be analyzed. These emerged from the goals of the inquiry which were to *improve the disputative and deliberative argument skill of my students*. The typologies identified for this analysis are (1) statements of strategic approach for argument construction, and (2) statements of knowledge regarding how to construct strong scientific arguments. Data included in the analysis were the responses students had to the interview questions only. No other utterances were analyzed. Sections of the interview transcripts were marked as belonging to one of the typologies or the other. Responses to questions that demanded argument construction were associated with the strategic approach typology, while responses to the final question in each interview, “What makes a strong argument?”, were related to knowledge about argument construction and were marked as being of the second typology.

Reading through the transcripts of the interviews several times, I subjectively felt that students addressed the questions in a manner that depended on many factors. These included, but were not limited to, extent of problem-solving experience, level of conceptual understanding of the general topic, heuristics that have been effective in the past, skill at using logic, and epistemic considerations that included methodological trust, all of which can play a role in argument construction. The term *strategic approach* is used here to indicate “what is going on?” in the mind of a student while constructing the argument response and to unearth features such as those listed above. However, they were not directly measured. Instead, they were hinted at in the student responses. This qualitative analysis attempted to gather those hints systematically, and draw conclusions based on insights that the data provided. Some portion of each student response contained utterances that were not associated with strategic approach such as (1) a restatement of the question, (2) internal rambling of thought, or (3) filler words like “um”, “so”, and “OK”. These were words that perhaps set up a thought, but weren’t needed to express it. This is not to say that any of the above did not play a role in helping the student respond to the question. Certainly, repeating the question may helped the student to mentally clarify it. But, since this analysis was trying to identify those statements that objectively played a role in the construction of an argument, restatement of the question to start a response did not move the student toward anything that could be objectively analyzed. The student was still at the starting point of the work and hadn’t measurably proceeded toward an answer.

Conversely, the students made many statements (or utterances) that indicated what they were thinking, at the moment, in relation to argument construction. These statements provided clues related to the strategy a student used to respond to the questions. For purposes of this analysis, statements were defined as being one or more sentences, or sentence parts, that made one point in the argument. For example, “Those aren’t fitting the trend” or “Density is mass over volume.”

Based on a thorough review of the transcripts, it was hypothesized that statements that indicated strategic approach could be categorized as (1) those that “declared” and (2) those that “made use of…”. Importantly, no judgement needed to be made as to whether the statements were true when categorized in this way. The list below describes these statement categories:

1. <u>Statements that declared a fact:</u> The student stated a given piece of information such as, *“Object 1 is red.”*

2. <u>Statements that declared a concept</u>: The student stated a scientific idea such as, *“Density is mass over volume.”*

3. Statements that declared a procedure: The student stated either what they were doing or what they were planning to do. For example, *“I'd say the first thing that I look at is the temperature that they're all sitting at right now.”*

4. Statements that made use of facts (without use of science concepts): The student stated correlations, comparisons, trends, or other patterns in the given information such as, *“There is no correlation between the color and the density data”* or *“Object A undergoes a greater temperature change than object B.”*

5. Statements that made use of science concepts (with or without specific facts): The student used a scientific idea to make an assertion. Here is an example with fact usage: *“If the mass is one and the density is ten, then the volume is ten.”* Here is an example without fact usage: *“I can find the volume if I know the mass and density.”*

Students also made statements that indicated their understanding and knowledge of effective argument structure that fit the second typology. These were associated with responses to the final question of the interview protocol. Some of these statements identified *citing evidence* as important when constructing strong scientific arguments. For example, “A strong scientific argument is going to be based in fact.” Other statements pointed to the importance of *constructing a coherent argument*, as is seen in, “…and then can be easily followed by the reader in a progressive flow, so that the starting facts can then be connected all the way down.” Finally, other statements indicated that *conceptual rationale*, linking the evidence to the claim, was important as in, “And um, it’s supported by research. Um, I would also say that it’s in line with prior research unless it’s seeking to contradict that.”

Once the transcripts were coded to identify the various statement-types, patterns (regularities), relationships (links), and themes (integrating concepts) were identified and used to help answer the inquiry questions. The term *coverage percent* was used to identified what percentage of student utterances fell into the various statement-type categories. This concept will be described in greater detail in chapter four.

3.5.4 Use of Second Coders

A codebook was written so that evaluation of data could be done accurately and with consistency by me and by the second coders I recruited to help with the study. The codebook contained the rubrics, definitions, examples, and notes that were referred to when making the evaluations. The sections of the codebook that were pertinent to each of the second coders were shared and discussed. Whenever evaluation differences occurred, the second coder and I discussed each case and decided upon a consensus evaluation. At the conclusion of analysis point four, confidence in the evaluations reached a level (consistently above 80%) wherein it was no longer deemed necessary to continue double coding. The consensus evaluations, those that were agreed upon without discussion, and those that were not double-coded, were included in the data set used in the analysis of results in chapter four. An overview of the double coding is presented in Table 12.

Table 12. Second coder data

Data Category	Rubric	Second Coder Number	Percent Double Coded	Average Percent Agreement
Pretest (Quantitative)	Structure	1	26%	89%
Pre-Interview (Quantitative)	Structure	1	33%	80%

Pre-Interview (Qualitative)	Strategic Approach	2	22%	86%
Pre-Interview (Qualitative)	Knowledge	2	22%	100%
Lab Argument (Quantitative)	Structure and Accuracy	1	36% (Average over 4 units)	87% (Average over 4 units)
False Claim Response (Quantitative)	Structure and Accuracy	1	40% (Average over 4 units)	86% (Average over 4 units)
Free-Response (Quantitative)	Points-Based	3	29% (Average over 4 units)	83% (Average over 4 units)

4.0 Results

The data presented in this chapter provide evidence that will be used to answer the inquiry questions of this investigation. Each of the three sections that follow address one of the questions. Section 4.1 displays and discusses both quantitative and qualitative data that allowed me to determine the incoming argument skill level of my students. Section 4.2 focuses on student-created artifact evaluations that are quantitative in nature and were evaluated over the entire length of the study. In this section I will describe the changes in argument skill measured through unit five of the course. Section 4.3 presents and discusses quantitative and qualitative data that help to judge the argument skill level of my students at the conclusion of the study relative to precourse levels and to the standards set by the College Board.

4.1 Incoming Argumentation Skill Level

Inquiry question one asked, “What is the incoming level of argumentation skill of my students?” This was determined in two ways. First, by administering a pretest designed to assess skill in constructing a basic scientific argument that contained a claim, evidence and rationale. Second, through precourse cognitive appraisal interviews with a subsample of participants intended to provide insight related to argument strategies and knowledge.

As will be discussed in the following two subsections, these assessments supported the conclusion that while some students were able to construct full scientific arguments in response to precourse prompts, all incoming students needed to improve their skill level so that they could effectively respond to argument-related questions on the AP Physics 1 exam.

4.1.1 Precourse Argument Construction

The evidence presented here supports a key finding: precourse argument constructions almost always included some amount of evidence but often lacked explicit conceptual justification. I will first discuss data associated with student use of evidence and then turn my attention to that related to conceptual justification.

Individual pretest results (along with group averages, standard deviations, and relative standard deviations) are shown in Table 13, ranked according to average student structure evaluation for all three questions. Overall results from the pretest are shown graphically in Figure 6. Evaluations were made using the structure rubric found in Table 7 (p. 64) wherein arguments were judged on a scale of one to four indicating the inclusion of various argument components. A level one evaluation simply required an unsupported claim while a level four evaluation indicated a complete argument construction. A level two response required the use of evidence, but a level three response indicated that the evidence was sufficient and appropriate. Explicit use of conceptual justification was required for a level four evaluation. The entire pretest can be seen in Appendix A. All study participants completed this online assessment.

Table 13. Individual pretest results

Student	Rank	Pre Q1	Pre Q2	Pre Q3	Average
134	1	4	4	3	3.67
682	**1**	**4**	**3**	**4**	**3.67**
948	1	4	4	3	3.67
172	4	4	3	3	3.33

481	4	2	4	4	3.33
531	4	4	3	3	3.33
860	**4**	**2**	**4**	**4**	**3.33**
879	4	4	4	2	3.33
275	**9**	**2**	**4**	**3**	**3.00**
992	**9**	**4**	**2**	**3**	**3.00**
414	11	2	3	3	2.67
689	**11**	**2**	**3**	**3**	**2.67**
214	13	2	2	3	2.33
886	13	2	2	3	2.33
441	**15**	**2**	**1**	**2**	**1.67**
M		2.93	3.07	3.07	3.02
SD		1.03	0.96	0.59	0.58
RSD		34%	32%	19%	19%

Notes. *RSD* is relative standard deviation. Students in bold font were also interviewed prior to the start of the course

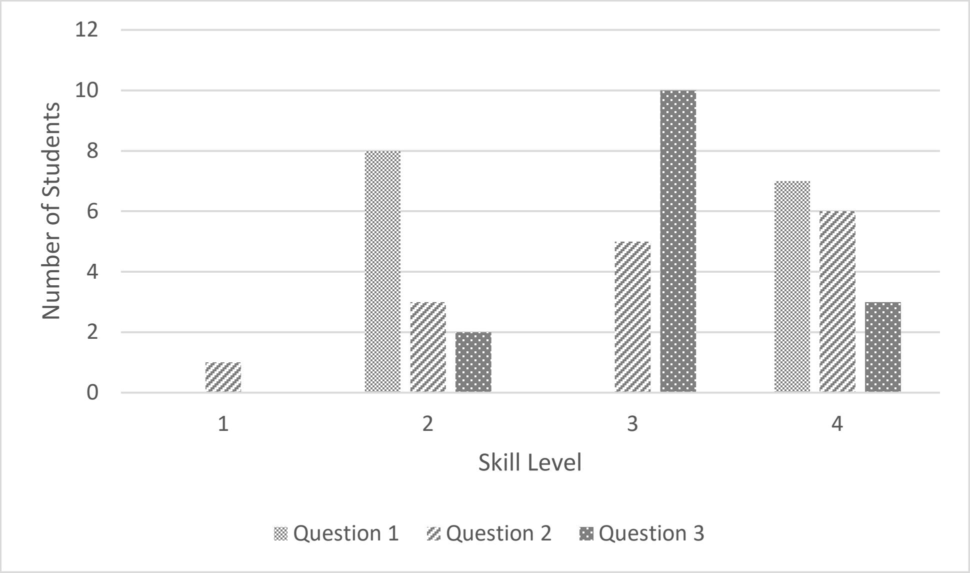


Figure 6. Pretest overall results

All but one pretest response contained some amount of evidence, allowing the average evaluation to be 3.02 over the entire assessment. So, it can be concluded that, in general, students understood the importance of including evidence in arguments constructed on the pretest. The results of the pretest were used to randomly select students to be interviewed prior to the course from within three sections of the participant group (low, middle, and high performers) so that a reasonable cross-section of students at various pretest skill levels could be further investigated. For a detailed look at the interview protocol, see Appendix B. Table 14 presents argument structure data on the individual interviewees.

Table 14. Quantitative data for structure evaluation - precourse interviewees

Student Number	Pretest Rank	Pretest Average	Question 1 Claim 1 TC	Question 1 Claim 2 D	Question 1 Claim 3 TC	Question 1 Claim 4 TC + D	Question 2 D	Precourse Interview Average
682	1	3.67	2	4	3	3	4	3.20
860	4	3.33	3	2	3	2	2	2.40
275	9	3.00	2	4	3	3	3	3.00
992	9	3.00	3	4	4	2	4	3.40
689	11	2.67	3	4	2	4	4	3.40
441	15	1.67	2	2	3	3	4	2.80
M		3.02	2.50	3.33	3.00	2.83	3.50	3.03

Note. TC: question related to thermal conductivity; D: question related to density

There were no precourse interview responses that did not include at least some evidence. The average score was approximately three (*M* = 3.03). Again, since the structure rubric demanded sufficient and appropriate evidence to evaluate an argument at level three, it is apparent that the importance of evidence in argument construction was generally understood by the interview group as a whole, but the quality of evidence presented was not uniform. Nearly a third of the responses utilized evidence that was either insufficient or inappropriate. This finding is in line with results from the pretest (*M* = 3.02), and also indicates that many (two-thirds) of the precourse assessment responses lacked the explicit conceptual justification needed to construct a full argument.

Of the forty-five pretest responses (Table 13, p. 83), 64% did not include statements of conceptual rationale. Similarly, twenty of the thirty (67%) interview arguments lacked explicit conceptual justification (Table 14, p. 85). Without this argument feature a level four evaluation was not possible. Assuming that the students wanted to do their best on the precourse assessments, the reasons for these findings may have been two-fold. First, the specific conceptual understanding required to justify evidence in an argument may have been lacking, depending upon the topic of the prompt. This can be inferred from the average evaluation of the interviewees for prompts related to thermal conductivity that were always less than those associated with density. Second, there may have been lack of knowledge, at the group level, that conceptual justification is an important part of argument construction. I will now address this second factor in some detail by outlining the findings of a qualitative analysis performed on the precourse interview data.

In order to identify the strategic approach used by the precourse interviewees to construct arguments, transcript data were analyzed using a measure called "coverage percent." This term refers to the extent to which a student response is taken up by specific types of statements relative to the entire interview (see Appendix B) and was calculated by the NVivo software application used in the analysis in response to queries on the data set. These statement types included declaration of fact, declaration of concept, declaration of procedure, use of fact, use of concept, and "other". Statements in the "other" category included such utterances as a restatement of the prompt and non-sensical rambling that did not contribute to the argument construction in a way that could be objectively determined. In this analysis, I used coverage percent as a proxy for the answer to the question "what is going on?" in the student's cognitive process when constructing an argument. In other words, what is being cognitively generated in a strategic manner? For example, student 275 uttered the following as *part* of the response to claim four of question one. This statement was coded in the “use of fact” category and represented 2.76% of the entire interview for that student.

> “I'm going to have to agree with the statement because the two objects with the greatest density gained the most heat by a significant factor. So, the statement is true.”

Notice that the student did not simply declare the fact that the two objects mentioned had specific density values or that they gained a specific amount of heat. Instead, a comparison was made in reference to other objects discussed in the prompt. The facts *were used* to make the comparisons.

The data in Table 15 provide an overview of the interview data expressed in terms of the *average* coverage percent, broken down by question and student structure evaluation. The coverage percent values shown are not for any single student. Instead they are averages for all interviewees who were evaluated at specific structure levels for each prompt.

Table 15. Average coverage percent by question, statement type, and structure level

	Declare Fact			Declare Concept			Use of Fact			Use of Concept		
Structure Level =>	2	3	4	2	3	4	2	3	4	2	3	4
Q1 C1	4.54	7.66		3.71	7.49		34.90	18.90		28.09	33.78	
Q1 C2	9.51		4.40	3.69		6.56	43.20		18.75	19.54		36.63
Q1 C3	3.75	7.13	4.35	4.23	4.10	12.97	13.04	32.87	16.89	57.18	26.19	23.64
Q1 C4	9.62	4.54	3.75	9.12	3.71	4.23	21.83	34.90	13.04	22.08	28.08	57.18
Q2	14.88	3.41	4.58	5.27	3.99	6.09	26.76	30.71	25.98	20.52	27.24	34.46
M	**8.46**	**5.68**	**4.27**	**5.20**	**4.82**	**7.46**	**27.95**	**29.35**	**18.67**	**29.48**	**28.82**	**37.98**

The total coverage percent for these four statement types does not add up to 100% because there were other utterances, such as procedural statements, rambling, and statements made in response to the final (non-argument construction) question, that account for the remaining coverage. Since there were four parts (claims) to question one, they are shown separately as C1 to C4. Here is an example of how to read the data in Table 15: students who were evaluated at level three for their response to claim one of question one, on average, used 7.66% of the response *declaring* a fact. Whereas, those same students used, on average, 18.90% of the response *using* a fact. Blank fields in the table indicate that none of the six interviewees were evaluated at that level for that particular question.

A careful review of this data led to interesting findings. Students who constructed arguments evaluated at level four declared and used concepts more than other students. Also, students who constructed arguments at levels two and three, declared and used facts more than students evaluated at level four. This may mean that students who rely on facts tend to make less robust arguments than those who rely on concepts.

Generally, making use of facts relies on logic and critical thinking (comparisons or correlations, for example) to rationalize the use of evidence presented in an argument, while making use of concepts relies on theoretical understanding of science principles to connect evidence to a claim. For example, when students cite factual "trends or patterns" in data that are assumed to be continuous, they are supporting their argument in a way that goes beyond simply presenting evidence, but they aren't necessarily making a very strong case. Using this type of rationale is still a good thing to do, as recognition of the need for extra support is indicative of the understanding that a robust argument needs to have more than just a claim and evidence. However, logical rationale using facts may not always *solidly* support a link between the claim and the evidence presented in an argument. In order for the argument to be strong, conceptual support is needed. Here is an example of an argument from one of the interviews (student 689) that makes this point:

> "It seems as though there's not a correlation. That there's two brown objects – one of them heats up quickly and one of them heats up slowly. And there's three light objects…and…it looks like the average would be similar to the brown. So, I would say disagree."

The weakness of an argument with this type of rationale is that the limited amount of data may not be representative of the way nature works in general. Note how much stronger the argument would be if the following underlined conceptual justification was added:

> "It seems as though there's not a correlation. That there's two brown objects – one of them heats up quickly and one of them heats up slowly. And there's three light objects…and…it looks like the average would be similar to the brown. Theoretically, the rate at which an object changes temperature is controlled by the microscopic structure of the object. I don't think that color depends upon microscopic structure in the same way. This makes sense relative to the lack of correlation found in the data. So, I would say disagree."

This is not to say that logical justification using facts is unimportant to a scientific argument. It often plays a pivotal role in the argumentative process, connecting a conclusion to the given factual premises syllogistically. For example, this excerpt from student 682 shows how logical comparison can be a key part of an argument:

> "Object F is denser than object C. But it is equally dense as object D. Object D's density is four and it also floats. Therefore, object F with a density of four, would float in water."

Responses to question three in the interview protocol also pointed to a lack of understanding of the importance of conceptual justification in argument construction. That question asked students to state what they believed made an argument strong. Coverage percent statistics for that part of the interview, presented in Table 16, indicate that students described the importance of evidence and coherency, on average, almost eight times as extensively as the importance of conceptual justification.

Table 16. Final question coverage percent by category

Aspects of a Strong Argument Stated by Student Interviewees	Average Coverage Percent
Importance of Coherency	3.7%
Importance of Evidence	3.9%
Importance of Conceptual Justification	0.5%

Here is a typical response to the final interview question from student 682. Note the emphasis on evidence and logic with no explicit mention of a conceptual rationale.

> “I think you first have to take note of observations that you make. So, if you are looking for certain information then you should – once you find it – you should point it out. Um, so like if I were looking for like the mass of something or the ratio of something, I would say that this object has this density with whatever unit. Um, and I think after you state all the things that you know and then you can logically create a logical argument based on information that you know. So, you shouldn’t just jump straight to an argument. You should first lay out everything that you already know and will use.”

After making specific reference to observational evidence, the student makes the case that connections should be made based on logic. It’s clear that the student recognizes the importance of the use of facts in effective argumentation, but there is no indication that justification of evidence should be conceptually based.

4.1.2 Inconsistent Exhibition of Precourse Argumentation Skill

A second key finding of the precourse assessments was that students did not perform consistently when constructing arguments. I will first discuss the quantitative pretest data followed by both quantitative and qualitative precourse interview data that support this finding.

As is shown in Table 13 (p. 83) of the previous subsection of this report, not one of the fifteen students had the same score on all three questions of the pretest assessment. There was also variability in group performance on each question as measured by the relative standard deviation (*RSD*). Question one had the lowest average score (2.93) and the highest *RSD* (34%). Questions two and three showed identical average evaluations (3.07), but question two (*RSD* = 32%) had greater evaluation variability than question three (*RSD* = 19%). I inferred from these results that although responses to the pretest questions generally contained some amount of evidence, the quality and quantity of the evidence was not consistently presented at either the individual level or the group level.

Inconsistent performance was also noted on the precourse interviews. Table 17 presents data on the six interviewees individually. Data associated with all three interview questions are included, but average coverage percentages for questions one and two are presented together, as they both demanded argument construction while the final question did not. Students are listed in order of their pretest performance rank.

Table 17. Individual precourse interview coverage percent data

				Questions 1 and 2						
Student number	Pretest Rank	Pretest Average	Precourse Interview Average	Declare Fact %	Declare Concept %	Use Fact %	Use Concept %	Declare Procedure %	Other %	Final Question
682	1	3.67	3.20	6.08	5.03	14.36	38.45	9.62	11.69	14.77
860	4	3.33	2.40	14.88	5.27	26.76	20.52	0.00	20.20	12.37
275	9	3.00	3.00	3.41	3.99	30.71	27.24	19.80	7.38	7.38

992	9	3.00	3.40	4.35	12.97	16.89	23.64	15.77	23.40	23.40
689	11	2.67	3.40	3.75	4.23	13.04	57.18	5.58	3.35	3.35
441	15	1.67	2.80	4.13	2.11	59.64	18.56	0.00	6.18	6.18
M		2.89	3.03	6.10	5.60	26.90	30.93	8.46	12.03	11.24
SD				4.40	3.78	17.53	14.65	8.18	8.09	
RSD				72%	67%	65%	47%	97%	67%	

Two insights are noteworthy. The first is that even though it’s clear that the students who declared and used concepts a relatively high percentage of the time (students 992, 689, and 682) must have extensively used them in argument construction, that doesn’t mean they used them to justify *each* argument they constructed (Table 14, p. 85). Conversely, students who declared and used facts a high percentage of the time, still sometimes used concepts to justify an argument. The point is that students used their understanding of how to construct an argument differently in different situations. Possible explanations for this lack of consistency will be addressed in chapter five.

Here are two examples from student 275 that exemplify this point. The first is in response to interview question one, claim two. The second is in response to interview question 1, claim three. To make the evaluation easy to understand, claims are shown in bold font, evidence is italicized, and conceptual justification (only found in the first example) is underlined.

> Example 1 (evaluated at level 4): “Okay well. It's gonna be very simple that this is going to be a mass and density question. So, we know that density equals mass times volume and then because we said, recall the following - the size of a sphere is directly related to its radius. So, the volume and the radius - if they have the same volume, they’re gonna have the same radius. So, the density equals mass times volume - wait - mass over volume. Density equals mass over volume. I think...So, you can plug in your different objects to see what their volume is going to be based on their density and mass. So, *object D will have a density of* - *er* - *a volume of two.* I'm plugging in and solving for the missing variable in D

equals M over V. *C will have a volume of one. D will have a volume of two. We know that A has a very statistical anomaly, um. It's a very different density.* So, I think we can rule that one out. *And object B also, the same is true*. So, *objects B and object C have the same volume.* **So, I would agree with that statement.**"

Example 2 (evaluated at level 3): "Now we're just going to have to also take into the density column, as well as the temperature change column. *And object A has a density of 10 and it gains of the most heat when placed in the oven. And object E had a density of 15 and gained the second most.* So, if you look at the - okay well *objects B, C and D don't seem to fit the same trend because objects B and D have a greater density - gained less heat than object C. So, those aren't fitting the data trend*. There is something else that tells us. I'm going to have to agree with the statement because *the two objects with the greatest density gained the most heat by a significant factor*. **So, the statement is true**."

The first example is a complete argument. The student used concepts to justify the use of specific evidence. For example, the volume was not given, but the concept of density was used to find it, adding to the pool of available evidence. Finally, the concept of volume was related to the size of the object, allowing the evidence provided to connect to the claim. The second argument does not connect the evidence to the claim in a conceptual way. Instead, the student used an observed trend in the data to draw a conclusion, assuming that the relationship pointed to in the trend was more than coincidental. Adding a conceptual justification would strengthen the argument by giving the reader a reason to think that the identified trend is important and that it is linked to the claim by more than logic alone.

A second insight regarding inconsistency is that individual average pretest evaluation was not predictive of performance on the precourse interview questions. Table 17 (p. 91) shows a side-by-side comparison of average structure scores for the interviewees on the pretest and the precourse interview. Results of the Pearson correlation computation indicated that there was no association between the results of these two assessments, (*r*(4) = .069, *p* = .897).

4.2 In-Course Patterns of Progress

Inquiry question two asked, "What patterns of progress in argumentation skill development do students exhibit as they engage in the constellation of argumentation activities offered at Physics Prep?" Three types of in-course assignments required students to construct arguments whose evaluations were used to answer this question. These were, (1) an argument embedded in a lab report, (2) a rebuttal to a teacher-created false claim (FC), and (3) a paragraph-length argument construction in a free-response (FR) question on a unit test. Each assignment type was evaluated in units one through five using the structure rubric (Table 7, p. 64) and the accuracy rubric (Table 8, p. 68). Appendix C lists all the prompts for each assignment.

Artifacts collected throughout the investigation indicated that students did, as a group, improve their skills up to a certain level and retained them. The general pattern of progress showed an initial positive slope that flattened from unit three to unit five. However, when analyzed individually, students did not consistently exhibit their ability to construct arguments to the same level for all assignment types. One of the insights that became apparent through data analysis was the importance of physics-related conceptual understanding of the course material as a foundation for argument construction. The next two subsections will present evidence to support these conclusions.

4.2.1 In-course Argument Construction

The evidence presented in this subsection supports another key finding: student argument construction improved at the group level over the course of the investigation. I will support this finding through an in-depth discussion of the structure evaluations of student-constructed artifacts submitted in units one through five. In line with this finding, I will also discuss the relationship between constructing a complete argument and the accuracy of its components. It is through this comparison, along with other evidence presented in this report, that the importance of physics-based conceptual understanding will be discussed in chapter five.

Every student artifact collected in this investigation was evaluated for structural quality and accuracy. Table 18 presents the structure evaluations for argument constructions from all five units. Submission rate data is included along with descriptive statistics for each assignment. Average evaluation scores for all three assessments associated with a single unit are also shown.

Table 18. Structure evaluations for artifacts from all five units

Student ID	Pre-Test Rank	Lab 1	FC 1	FR 1	Lab 2	FC 2	FR 2	Lab 3	FC 3	FR 3	Lab 4	FC 4	FR 4	Lab 5	FC 5	FR 5
134	1	**4**	**2**	**4**	**4**	**4**	**3**	**4**	**4**	**NC**	**2**	**4**	**4**	**2**	**4**	**4**
682	1	**2**	**4**	**4**	**X**	**3**	**4**	**4**	**2**	**NC**	**4**	**4**	**1**	**4**	**4**	**4**
948	1	**2**	**2**	**1**	**4**	**3**	**2**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
172	4	**3**	**1**	**4**	**X**	**4**	**3**	**X**	**4**	**NC**	**X**	**1**	**2**	**4**	**2**	**2**
481	4	**3**	**2**	**4**	**4**	**4**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
531	4	**4**	**2**	**4**	**4**	**4**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
860	4	**2**	**2**	**2**	**2**	**4**	**2**	**3**	**2**	**4**	**4**	**4**	**NC**	**4**	**3**	**3**
879	4	**2**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**2**	**4**	**4**	**4**	**4**	**4**	**4**
275	9	**4**	**4**	**4**	**3**	**4**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
992	9	**4**	**4**	**4**	**3**	**2**	**4**	**2**	**4**	**4**	**X**	**4**	**2**	**3**	**2**	**2**

414	11	**4**	**2**	**4**	**2**	**4**	**3**	**2**	**2**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
689	11	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
214	13	**2**	**4**	**4**	**4**	**4**	**X**	**X**	**4**	**NC**	**4**	**4**	**2**	**X**	**4**	**4**
886	13	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
441	15	**1**	**2**	**2**	**4**	**4**	**3**	**4**	**2**	**4**	**4**	**4**	**4**	**3**	**2**	**4**
Not Submitted		0	0	0	2	0	1	2	0	0	2	0	0	1	0	0
Submitted		15	15	15	13	15	14	13	15	15	13	15	15	14	15	15
% Submitted		100	100	100	86.7	100	93.3	86.7	100	100	86.7	100	100	93.3	93.8	100
M		3.00	2.87	3.53	3.54	3.73	3.21	3.62	3.47	3.82	3.85	3.80	3.36	3.71	3.53	3.67
SD		1.07	1.13	0.99	0.78	0.59	0.70	0.77	0.92	0.6	0.55	0.77	1.08	0.61	0.83	0.72
Unit Average		1	=	3.13	2	=	3.51	3	=	3.64	4	=	3.61	5	=	3.64

Notes. X is not submitted; NC is no claim; Darker shading indicates a higher evaluation.

Overall, 96% of assignments were submitted for evaluation. However, no evaluation was done on the five FR submissions that were lacking a claim or were left blank (NC). Thus, 94% of submissions were evaluated. Group unit averages (bottom row) are based on the average unit score for each student so that all students carry the same weight per unit even if one of the three submissions in a particular unit was marked NC or X.

There are various ways to tease out the meaning of the data contained in the artifact overview table (Table 18), but one visual way to consider its implications is to see how the shading of the structure evaluations become generally darker from one unit to unit five. The darker the shading, the higher the evaluation. This doesn’t mean that individual students consistently improved when constructing arguments. Instead, it seems apparent that, as a group, students improved their argument construction skill relative to that measured at the start of the course. This interpretation is supported by the scatter-plot shown in Figure 7 that shows the group structure averages for the pretest and for each unit.

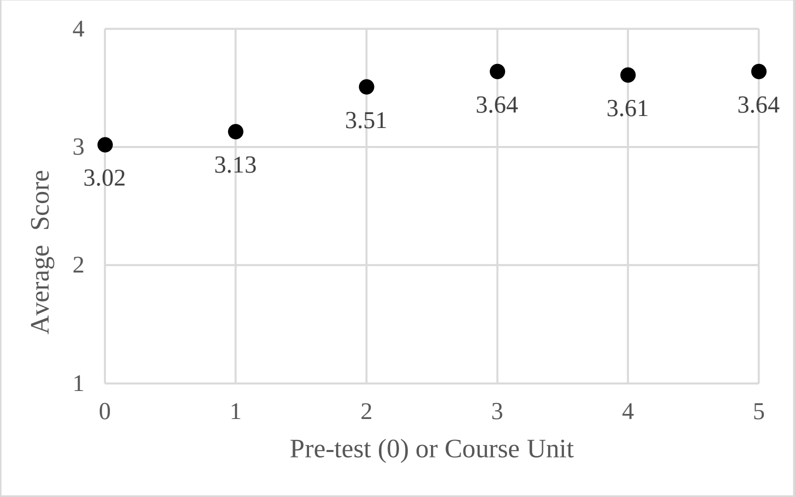


Figure 7. Class structure average by unit

The data indicate that after a small gain in unit one (*M* = 3.13) relative to the pretest assessment (*M* = 3.02), larger gains in units two and three were maintained through unit five. These results were tested for effect size using paired-sample T-tests (two-tailed) comparing pretest results with average structure scores for every student (*N* = 15) in each unit. Table 19 shows effect size (using Cohen's d) for various combinations of analysis points (pretest and units).

Table 19. Effect size table (Cohen's d) for student artifacts by analysis point

	Compared with paired-samples from…				
Student Average from…	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5
Pretest	0.10952	0.59181	0.84810	0.58127	0.79207
Unit 1		0.50396	0.59386	0.44361	0.57482
Unit 2			0.29017	0.12222	0.22649
Unit 3				0.03752	0.00032
Unit 4					0.05753

Note. Small Effect Size > 0.2; Medium Effect Size > 0.5; Large Effect Size > 0.8

The shading in Table 19 is indicative of the effect size. The darker the shading the greater the level of effect. I am interpreting the effect size as simply indicating how much more effective the students were at constructing a scientific argument at one point in the investigation compared to another. This measure does not indicate statistical significance. When compared with average pretest score (*M* = 3.02, *SD* = 0.58) a medium effect size was measured once students completed work in unit two (*M* = 3.51, *SD* = 0.43), *t*(14) = 2.29, *p* = 0.038, *d* = 0.59181. Thereafter, medium and large effects were measured, relative to the pretest, for the remainder of the study. There were also small and medium effect sizes for other unit comparisons, but the relative stability of the results from the end of unit two until the completion of the unit five is noteworthy and will be discussed in chapter five.

Accuracy data were also gathered on every student-created artifact in units one through five using the rubric shown in Table 8 (p. 68). Table 20 presents that data for the entire investigation in a similar way that structural data was presented in Table 18 (p. 95). Again, no evaluation was made for FR submissions that had no claim (NC). As before, darker shading indicates a higher evaluation.

Table 20. Accuracy evaluations for artifacts from all five units

Student ID	Pretest Rank	Lab 1	FC 1	FR 1	Lab 2	FC 2	FR 2	Lab 3	FC 3	FR 3	Lab 4	FC 4	FR 4	Lab 5	FC 5	FR 5
134	1	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**NC**	**1**	**2**	**1**	**2**	**2**	**2**
682	1	**1**	**2**	**2**	**X**	**1**	**2**	**2**	**1**	**NC**	**2**	**2**	**0**	**1**	**1**	**2**
948	1	**2**	**1**	**2**	**2**	**2**	**1**	**2**	**2**	**1**	**2**	**2**	**1**	**2**	**2**	**0**
172	4	**2**	**2**	**2**	**X**	**2**	**2**	**X**	**2**	**NC**	**X**	**1**	**2**	**2**	**0**	**0**
481	4	**1**	**1**	**2**	**2**	**2**	**2**	**1**	**2**	**2**	**2**	**1**	**1**	**2**	**2**	**2**
531	4	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**1**	**2**	**2**	**2**	**2**
860	4	**1**	**0**	**1**	**1**	**2**	**2**	**1**	**1**	**2**	**1**	**1**	**NC**	**2**	**2**	**0**
879	4	**1**	**2**	**2**	**1**	**2**	**1**	**2**	**2**	**1**	**2**	**2**	**1**	**2**	**2**	**2**
275	9	**1**	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**1**	**2**	**2**	**2**
992	9	**2**	**2**	**1**	**2**	**1**	**2**	**2**	**1**	**1**	**X**	**2**	**1**	**1**	**1**	**2**
414	11	**1**	**2**	**2**	**1**	**2**	**2**	**1**	**0**	**1**	**2**	**2**	**1**	**2**	**2**	**0**
689	11	**2**	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**1**	**2**	**2**	**2**
214	13	**2**	**2**	**2**	**1**	**2**	**X**	**X**	**2**	**NC**	**2**	**2**	**1**	**X**	**2**	**2**
886	13	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**1**	**2**	**2**	**2**

441	15	**2**	**2**	**1**	**2**	**2**	**2**	**2**	**1**	**1**	**2**	**2**	**1**	**2**	**1**	**1**
Not Submitted		0	0	0	2	0	1	2	0	0	2	0	0	1	0	0
Submitted		15	15	15	13	15	14	13	15	15	13	15	15	14	15	15
% Submitted		100	100	100	86.7	100	93.3	86.7	100	100	86.7	100	100	93.3	93.8	100
M		1.47	1.60	1.80	1.69	1.87	1.86	1.77	1.60	1.55	1.85	1.73	1.07	1.86	1.67	1.40
SD		0.52	0.63	0.41	0.48	0.35	0.36	0.44	0.63	0.52	0.38	0.46	0.47	0.36	0.62	0.91

Notes. X is not submitted; NC is no claim; Darker shading indicates higher evaluation.

These evaluations helped to determine if any relationship existed between the structural evaluation of the artifact and the accuracy of the claim and its accompanying supports. Except for the FR submissions, like the structure data in Table 18 (p.95), the shading seems to be generally darker in later units. Figures 8, 9, and 10, show average lab argument data, FC data, and FR data, respectively, versus unit number for both structure and accuracy averages. A look at these scatter plots shows that improvements and diminishments of these two measures are visually similar from unit to unit. Results of the Pearson correlation computation indicated that there was a strong positive association between lab structure score and lab accuracy score (*r*(3) = .977, *p* = .004). FC structure score and accuracy score also showed a strong positive association (*r*(3) = .682, *p* = .205). FR structure and accuracy scores showed a weak negative association (*r*(3) = -.127, *p* = .839).

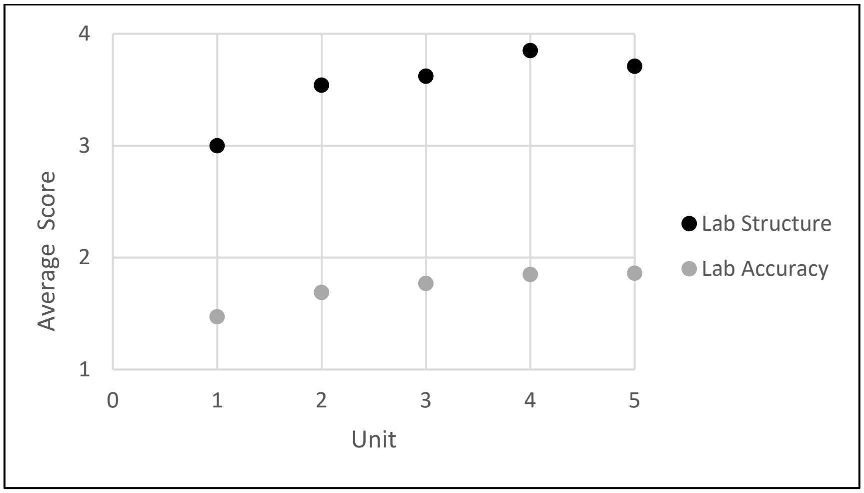


Figure 8. Class average for lab arguments by unit

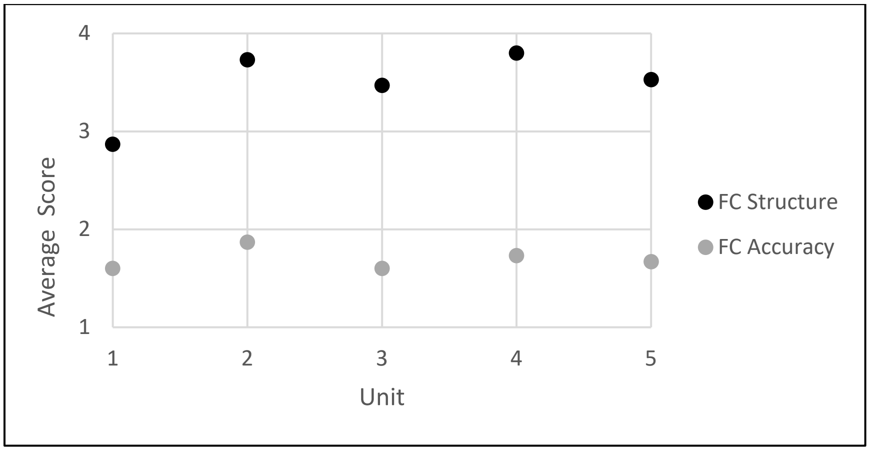


Figure 9. Class average for false claim responses by unit

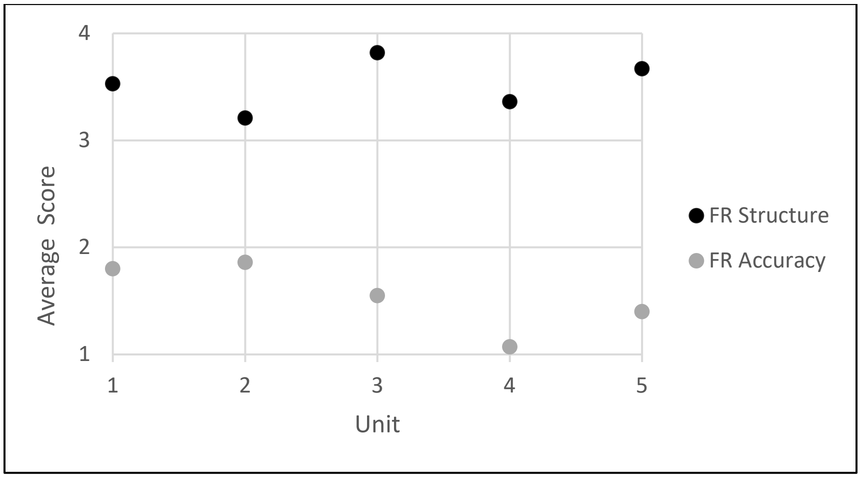


Figure 10. Class average for free-response arguments by unit

Additionally, overall average individual structure scores were compared to average individual accuracy scores for all five units. Results of the Pearson correlation computation indicated that there was a strong positive association (*r*(13) = .667, *p* = .006). Figure 11 shows a scatter plot of the data.

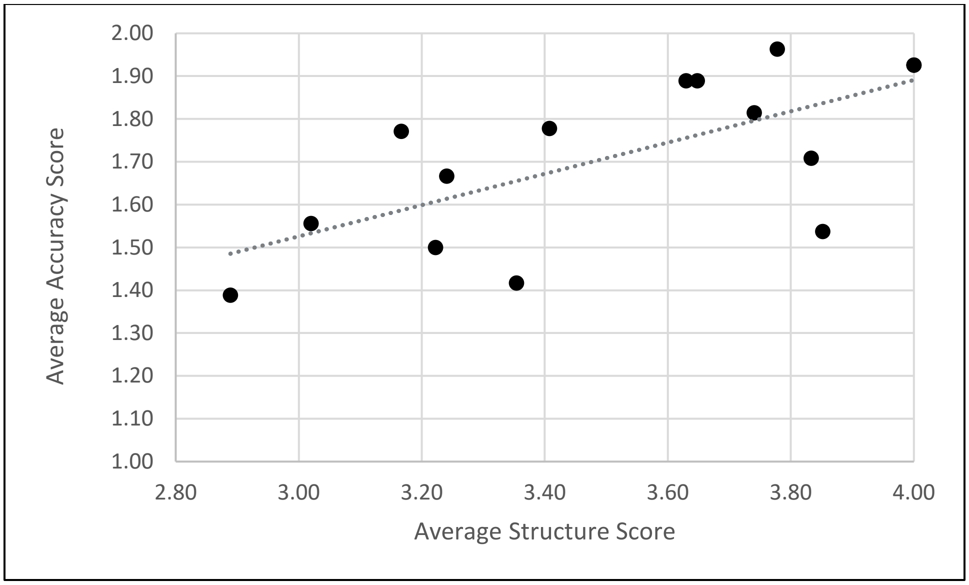
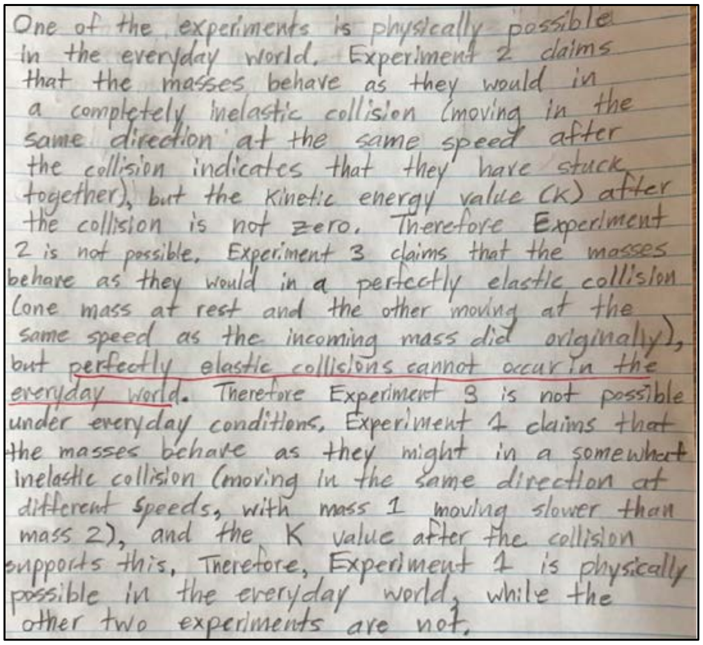


Figure 11. Average accuracy score vs. average structure score for individuals

The importance of the relationship pointed to in Figure 11 will be extensively discussed in chapter five. If it can be assumed that expressing an argument accurately is related to conceptual understanding of the topic associated with the prompt, these findings generally support the notion that success in constructing an argument depends upon conceptual understanding.

One might be concerned that the accuracy evaluation was merely measuring the same skill as the structure evaluation. However, care was taken when constructing the rubrics so that they evaluated two separate aspects of the argument. The structure rubric did look for sufficient and appropriate evidence, but the accuracy of the claim was not taken into account. More often than not, differences between structure and accuracy were noted when an accurate claim was made and some, but not all, of the evidence or rationale was inaccurate. Such an example (structure = 4; accuracy = 1) is shown in Figure 12.



One of the experiments is physically possible in the everyday world. Experiment 2 claims that the masses behave as they would in a completely inelastic collision (moving in the same direction at the same speed after the collision indicates that they have stuck together), but the kinetic energy value (K) after the collision is not zero. Therefore Experiment 2 is not possible. Experiment 3 claims that the masses behave as they would in a perfectly elastic collision (one mass at rest and the other moving at the same speed as the incoming mass did originaly), but perfectly elastic collisions cannot occur in the everyday world. Therefore Experiment 3 is not possible under everyday conditions. Experiment 1 claims that the masses behave as they might in a somewhat inelastic collision (moving in the same direction at different speeds, with mass 1 moving slower than mass 2), and the K value after the collision supports this. Therefore, Experiment 1 is physically possible in the everyday world, while the other two experiments are not.

Figure 12. Student example with high structure and mid-level accuracy evaluations

In this case, student 134 made an accurate claim in the first sentence, and included appropriate and sufficient evidence. However, one of the statements, "…but the kinetic energy (K) after the collision is not zero" is not a test for inelastic collisions as is implied. This inaccuracy does not however, lower the structure score for this response, as it contains plenty of other evidence and rationale that support the claim. The example given here is not meant to show that structure and accuracy are not correlated, but that there is no reason, based on the rubrics used, that they *must* be. A complete list of examples for possible combinations of structure and accuracy scores is shown in Appendix E.

4.2.2 Inconsistent Exhibition of In-Course Argumentation Skill

When analyzed on an individual basis, the evaluations presented in Table 18 (p. 95) of the previous subsection indicated that students did not perform consistently on all in-course assignments. Four of the fifteen students had evaluations ranging from level one to level four.

Eight students varied from level two to level four. One student earned only level three or level four evaluations. Only two were evaluated at level four for all assignments. This variation generally indicated that students improved over time. However, ten of the fifteen students exhibited evaluation drops at least twice *after* an evaluation of four was earned. This meant that many students who demonstrated the ability to construct a full argument subsequently failed to do so in later assignments.

4.3 Poststudy Argumentation Skill Level

Inquiry question three asked, “Are the resources offered at Physics Prep sufficiently robust to provide students the opportunity to reach levels of argumentation skill needed to effectively respond to assessment items they will encounter on the AP Physics 1 examination?” In order to answer this question, I analyzed student responses to FR argument-related questions on unit tests using point-based rubrics, posttest argument constructions using the structure rubric (Table 7, p. 64), and poststudy interview responses using the structure rubric and qualitative analysis. In section 4.3.1 I will discuss evidence drawn from the point-based rubrics, while in section 4.3.2 I will detail that produced in the poststudy assessments.

I believe that the evidence presented in the next two subsections indicates that course resources are adequate to improve student skill in argumentation to the needed level *as long as* that skill is coupled with necessary content knowledge.

4.3.1 Evidence from Free-Response Submissions Using Point-Based Rubrics

The evidence presented in this section supports two key findings used to answer inquiry question three. The first is that descriptive statistics of point-based rubric evaluations of FR argument constructions indicated general student success at these tasks. The second is that individual students did not perform consistently over the course of the investigation on FR argument construction.

As was described in section 3.4.1, the College Board develops new point-based rubrics to evaluate student performance on the FR section of the AP Physics 1 exam each year. Generally, for the paragraph-length argument question, the rubric awards points for content and structure (see Figure 4, p. 61). Therefore, rubrics used in this investigation to evaluate mastery of course material within a test question that required argument skill, were also point-based (see Appendix C). Results from the point-based evaluations are shown for each student in Table 21.

Table 21. Individual results for point-based free-response evaluations

Student ID	Pretest Rank	FR 1 (5 pts.)	FR 2 (6 pts.)	FR 3 (7 pts.)	FR 4 (7 pts.)	FR 5 (7 pts.)	*M*	*SD*	FR Rank
134	1	100%	67%	0%	86%	71%	65%	39%	10
682	1	80%	100%	14%	0%	71%	53%	44%	13
948	1	60%	33%	100%	71%	86%	70%	26%	9
172	4	100%	67%	0%	14%	43%	45%	40%	15
481	4	100%	83%	100%	71%	100%	91%	13%	5
531	4	100%	83%	100%	100%	71%	91%	13%	5
860	4	60%	67%	86%	14%	57%	57%	26%	12
879	4	100%	100%	100%	71%	100%	94%	13%	2
275	9	100%	83%	100%	86%	100%	94%	9%	2
992	9	100%	100%	86%	57%	43%	77%	26%	7
414	9	100%	83%	86%	43%	71%	77%	21%	7
689	9	100%	100%	100%	71%	100%	94%	13%	2
214	13	100%	X	29%	14%	100%	61%	46%	11
886	13	100%	100%	100%	86%	100%	97%	6%	1
441	15	20%	83%	57%	43%	43%	49%	23%	14
M		88%	82%	71%	55%	77%	74%		
Median		83%	86%	71%	71%	77%	85%		
Mode		100%	100%	100%	71%	100%	100%		
SD		23%	18%	38%	31%	22%	26%		

Note. X is not submitted

From the start of the course, it was apparent that, as a group, the students could effectively respond to the type of FR question that required a paragraph-length argument construction. The results showed that thirteen of the fifteen students earned full credit on at least one of the FR argument constructions by the end of unit three. Additionally, the descriptive group statistics shown at the bottom of the table indicated general success. Perhaps more importantly, the median value of all the evaluations was 85%, indicating that half of the responses were evaluated above this value, which is nearly fifteen points beyond what is needed to achieve the highest AP rank of extremely well qualified. Finally, the most common score, the mode of the data set, was 100%.

While these data suggest generally positive results, the range of FR scores for each student indicated that they did not perform in a consistent manner within this assignment type. Thirteen of the fifteen participants fluctuated three or more letter grades on their FR work over the five units of the investigation. However, students with the highest FR rank did show less variability as the top six students also had the six lowest standard deviations. Likewise, there was inconsistency demonstrated at the group level from unit to unit. Unit one had the highest average grade of 88% while unit 4 had the lowest at 55%. This also represents a range of three letter grades.

4.3.2 Evidence from the Poststudy Assessments

I will now discuss in detail the comparisons between pre- and post-assessments that help to answer inquiry question three. At the conclusion of unit five, all participants were given a posttest. The questions on the posttest are presented in Appendix A. Each question on the posttest was designed to mimic the associated pretest question. Student responses were evaluated using the structure rubric (Table 7, p. 64). The posttest results were compared to those from the pretest in several different ways. Table 22 shows the results as a side-by-side comparison.

Table 22. Comparison of pretest and posttest results

Student	Pre Q1	Post Q1	Pre Q2	Post Q2	Pre Q3	Post Q3	Pre-Ave	Post-Ave	Gain/Loss
134	4	4	4	4	3	4	3.67	4.00	0.33
682	4	2	3	4	4	2	3.67	2.67	-1.00
948	4	2	4	2	3	3	3.67	2.33	-1.33
172	4	2	3	4	3	4	3.33	3.33	0.00
481	2	2	4	4	4	3	3.33	3.00	-0.33
531	4	4	3	4	3	3	3.33	3.67	0.33
860	2	4	4	4	4	4	3.33	4.00	0.67
879	4	4	4	2	2	4	3.33	3.33	0.00
275	2	4	4	4	3	4	3.00	4.00	1.00
992	4	2	2	4	3	4	3.00	3.33	0.33
414	2	2	3	4	3	4	2.67	3.33	0.67
689	2	4	3	4	3	4	2.67	4.00	1.33
214	2	2	2	4	3	2	2.33	2.67	0.33
886	2	2	2	4	3	3	2.33	3.00	0.67
441	2	2	1	2	2	3	1.67	2.33	0.67
M	2.9	2.8	3.1	3.6	3.1	3.4	3.02	3.27	
SD	1.0	1.0	1.0	0.8	0.6	0.7	0.6	0.6	

Notes. Q is question; Dark shading indicates a gain; Light shading indicates a loss.

While the results from question 1 slightly decreased, those from questions two and three increased. Overall, the group average rose from 3.02 on the pretest to 3.27 on the posttest. Average evaluations for ten students showed an increase, two stayed steady, and three showed a loss. In all, forty-five responses were evaluated for both tests (three for each student). When looked at granularly, there were nine pre-post comparisons that showed a loss, twenty that showed a gain, and sixteen that remained the same. However, six of those sixteen were level four responses where no gain was possible. So, in all, only ten of the forty-five responses remained at level two or level three and didn’t improve. The gains primarily came from seven responses moving from level two to level four and from eleven responses moving from level three to level four. However, eight responses dropped from level four to a lower level, muting the positive results. The changes were not consistent for any one student as five showed a mixture of gain, loss, and unchanged evaluations, two showed evaluations that were either loss or unchanged, while seven had evaluations that were either gain or unchanged.

Another way to compare the pretest and posttest results is to look at the average scores for the group on each question of the assessments. Figure 13 presents these in a scatter-plot.

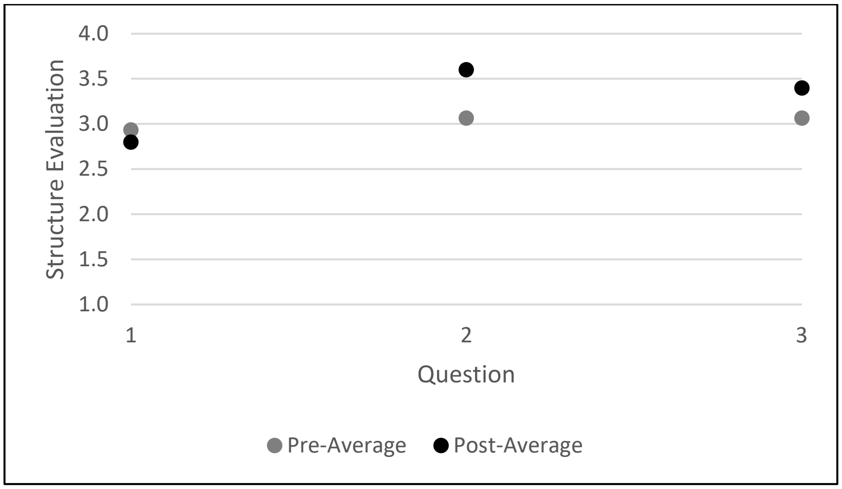


Figure 13. Pre/posttest structure comparison

The importance of the comparisons can be illuminated by measuring effect size using paired-sample T-tests. The results are presented in Table 23. I interpreted the effect size as measuring the difference in performance between pretest and posttest argument construction. The data were analyzed on a question by question basis for each student (*N* = 15) and by average score per student. The results indicated no effect for question one, small effects for questions two and three, and a small effect overall.

Table 23. Effect size information from paired T-tests pretest to posttest (Cohen’s d)

Pre/Post Comparison	*t*-test	*N*	*d*
Question 1 only	-0.367	15	0.09476
Question 2 only	1.658	15	0.42809
Question 3 only	1.234	15	0.31862
Average score per student (all three questions)	1.337	15	0.34521

Note. Small Effect Size > 0.2; Medium Effect Size > 0.5; Large Effect Size > 0.8

In addition to taking the posttest, the six students who were interviewed prior to the start of the course were interviewed again after the completion of unit five. Table 24 presents a side-by-side comparison of the structure evaluations for argument constructions from the precourse and poststudy interviews (see Appendix B for interview protocol details).

Table 24. Comparison of structure evaluations precourse interview to poststudy interview

SN	Q1 C1		Q1 C2		Q1 C3		Q1 C4		Q2		Q1&Q2		Q3	Q3	Q3
	TC		D		TC		TC&D		D		Ave.		C1	C2	Ave.
													M	D&E	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Post	Post	Post
682	2	3	4	3	3	3	3	3	4	4	3.20	3.20	4	3	3.50
860	3	2	2	3	3	3	2	3	2	2	2.40	2.60	2	2	2.00
275	2	3	4	4	3	4	3	2	3	4	3.00	3.40	4	4	4.00
992	3	4	4	4	4	4	2	4	4	4	3.40	4.00	4	4	4.00
689	3	3	4	4	2	2	4	3	4	4	3.40	3.20	4	3	3.50
441	2	2	2	4	3	3	3	3	4	4	2.80	3.20	4	3	3.50
M	2.50	2.83	3.33	3.67	3.00	3.17	2.83	3.00	3.50	3.67	**3.03**	**3.27**	3.67	3.17	3.42
SD	0.55	0.75	1.03	0.52	0.63	0.75	0.75	0.63	0.84	0.82	**0.39**	**0.45**	0.82	0.75	0.74

Notes. SN is student number; Q is question; C is Claim; TC is thermal conductivity; D is density; M is motion; E is energy; Dark shading indicates a gain; Light shading indicates a loss.

Where pre/post interview comparisons could be made, cells were shaded in light gray to show a loss and in dark gray to show a gain. Unchanged results were not shaded. In total, the poststudy interview protocol consisted of three questions that prompted students to evaluate seven different claims, and one question that asked to describe what makes an argument strong. Five of the seven claims were identical to those found in the precourse interview whose responses could be compared. Two claims were related to thermal conductivity, two were related to density, and one was related to both thermal conductivity and density. The final two claims were associated with motion and energy on question three and were not a part of the precourse interview protocol. The final question of the poststudy interview, wherein students were asked to give a description of a strong argument, was the same as that found in the precourse interview protocol and did not require argument construction.

Out of the thirty possible identical-prompt comparisons associated with questions one and two, four showed a loss and nine showed a gain. Importantly, eight of the seventeen that stayed the same were already at level four and could not show improvement. In all there were twenty responses that had the possibility of improvement, and of those, forty-five percent showed an increase. A paired-sample T-test between individual student average precourse evaluations (*M* = 3.03, *SD* = 0.39) and poststudy evaluations (*M* = 3.27, *SD* = 0.45), *t*(6) = -1.941, *p* = 0.110, *d* = 0.79259, indicated a medium level effect size for the six individual interviewees. However, when comparisons were made between average response evaluations *of the entire group* for the five prompts found in questions one and two, a paired-sample T-test, *t*(5) = -5.835, *p* = 0.004, *d* = 2.73083, found a large effect size. This is not surprising given the gains in average score for each prompt (see the next to the last row in Table 24, p. 108).

Examples of responses to one prompt (question one, claim two) from the precourse and poststudy interviews for student 441 are shown below. They are typical of those made by students who showed a gain in performance and are presented here to provide the reader with context. It should be noted that improvement was not measured for student 441 on any of the other four responses (see Table 24, p. 108). Claims are in bold font, evidence is italicized, and rationale is underlined. The interview protocols are presented in Appendix B.

Student 441, precourse interview question one (claim two):

> "*I'd say the objects yeah, the objects that are for example, if we take the objects B and D they both have - they have almost a similar temperature change when placed in the oven. They have a similar mass. Yeah, they have a similar mass. Yes, their*

> *densities are different. But they do float in water, both of them. And they both have a similar amount of warmth radiating from them.* **So, I agree that two of the objects are the same size.**"

Student 441, poststudy interview question one (claim two):

> "So, define whether they're the same size or radius you need to first find out the volume of the objects. So that's mass, so yeah density, since the density is given, the mass and the volume can be determined from that. Let me think one second. Can I have just one second to calculate all the? [short pause]. So, I calculated that the volumes of each of the objects and *then found that the volume of the object B and object C are equal because they are both mass M by D*. So, since the volume of the object, since there are spherical, should be 4 by 3 pi R cubed. So, since they're all supposed to – since they're all spherical, if the radius, er, If the volumes are equal that means the radii are equal, which means they are the same size. *So, in case of, since volume of B and volume of C are equal, both B and C will have the same radius and therefore be will be the same size*. **So, two objects are the same radius. So, I agree with the claim.**"

The first example was evaluated at level two. It does contain evidence, only some of which is pertinent to the prompt (asking if two of the objects had the same size). However, that evidence is insufficient to arrive at the stated conclusion. In contrast, the second example provides a fully constructed argument response to the same prompt. It contains sufficient and appropriate evidence along with conceptual justification that connects the evidence to the claim made in the final sentence.

As was done with data collected in the precourse interviews, a typological qualitative analysis was performed on data collected in poststudy interviews. Table 25 presents a side-by-side comparison of coverage percentages for each student for each statement type: declaration of fact, declaration of concept, use of fact, use of concept, declaration of procedure, and other.

Table 25. Comparison of coverage percent precourse interview to poststudy interview

			Pre-Questions 1 and 2 Post-Questions 1, 2, and 3												Final Question	
Student Number	Interview Average		Declare Fact %		Declare Concept %		Use Fact %		Use Concept %		Declare Proc. %		Other %			
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
682	3.20	3.20	6.08	6.07	5.03	6.94	14.36	39.71	38.45	23.47	9.62	7.29	11.69	8.17	14.77	8.35
860	2.40	2.60	14.88	6.56	5.27	17.18	26.76	35.24	20.52	12.90	0.00	0.00	20.20	23.54	12.37	4.58
275	3.00	3.40	3.41	5.79	3.99	15.04	30.71	16.26	27.24	40.62	19.80	11.15	7.38	3.90	7.47	7.24
992	3.40	4.00	4.35	4.55	12.97	9.47	16.89	15.95	23.64	36.06	15.77	14.46	23.40	11.23	2.98	8.28
689	3.40	3.20	3.75	8.78	4.23	16.74	13.04	5.25	57.18	34.21	5.58	13.43	3.35	8.72	12.87	12.87
441	2.80	3.20	4.13	8.00	2.11	7.70	59.64	30.57	18.56	32.59	0.00	3.46	6.18	11.37	9.38	6.31
M	3.03	3.27	6.10	6.62	5.60	12.18	26.90	23.83	30.93	29.97	8.46	8.30	12.03	11.16	9.97	7.94
SD	0.39	0.45	4.40	1.54	3.78	4.67	17.53	13.36	14.65	10.09	8.18	5.76	8.09	6.65	4.31	2.79
RSD	13%	14%	**72%**	**23%**	**67%**	**38%**	**65%**	**56%**	**47%**	**34%**	**97%**	**69%**	**67%**	**60%**	**43%**	**35%**

As a reminder, coverage percentage refers to the extent to which a student response is taken up by specific types of statements relative to the entire interview. Once again, as was found in the precourse interview data, the students showed variability as a group in their strategic approach to argument construction, as measured by the relative standard deviation (*RSD*) for each of the statement categories. For example, in the poststudy interview analysis, the use of concepts varied from a low of 12.90% for student 860 to a high of 40.62% for student 275. However, for each of the categories, the amount of variability decreased in the poststudy interview when compared to the precourse interview. Equally apparent was the inconsistent coverage percentage produced by each individual student when comparing precourse and poststudy interview results. For example, student 992 used concepts for 23.64% of the precourse interview but for 36.06% of the poststudy interview.

As a last analysis of the interviews, responses to the final question of the protocols were compared. That question asked students to state what they believed made an argument strong. Table 26 presents the results in terms of coverage percent.

Table 26. Comparison of coverage percent for the final question pre/post interviews

Aspects of a Strong Argument Stated by Student Interviewees	Average Coverage Percentage	
	Pre	Post
Importance of Coherency	3.7%	1.3%
Importance of Evidence	3.9%	2.1%
Importance of Conceptual Justification	0.5%	2.5%
Total	8.1%	5.9%
	(out of the 10.0% total coverage for response to this question)	(out of the 7.9% total coverage for response to this question)

Through qualitative analysis, statements were categorized into those that described the importance of coherency, evidence, and conceptual justification. Responses to this question also contained uncategorized statements that were placed into the “Other” category and are not represented above. Because students, in general, spent less of the poststudy interview responding to the final question, it was expected that the coherency and evidence coverage would be slightly less than for the precourse interview. But the reduction was more than might be expected given that only about 20% less coverage (10% compared with 7.9%) occurred for this question overall, while the coherency and evidence statements fell at a much larger rate (65% and 46% respectively). The large increase in interview coverage devoted to the importance of conceptual justification accounts for the change. There was a five-fold increase in the extent to which students mentioned concept-related rationale from precourse to poststudy for question three. Here are examples from the pre/post-interviews for student 682 that exhibit this change:

> (Precourse interview) “I think you first have to take note of observations that you make. So, if you are looking for certain information then you should – once you find it – you should point it out. Um, so like if I were looking for like the mass of something or the ratio of something, I would say that this object has this density with whatever unit. Um, and I think after you state all the things that you know and then you can logically create a logical argument based on information that you know. So, you shouldn’t just jump straight to an argument. You should first lay out everything that you already know and will use.”
>
> (Poststudy interview) “So, a strong scientific argument - you start with a claim first and then have to list, or have to give the facts surrounding the situation. So, you have to give us like the information basically to back up your argument and then you have to justify like why those different facts connect to each other, why they relate – like how, why they’re relevant basically to your argument. And then you have to, yeah you basically connect the facts through an argument. And also using scientific laws also helps. So, I’m running out of words. Yeah, you also have to wrap up a scientific argument. You can’t just leave it hanging and just...You have to put a conclusion so basically re-state your claim.”

Notice that in both examples the student begins by clearly stating the importance of evidence in argument construction. The precourse example then emphasizes coherent connection of the facts, based on logic, to form a conclusion. No mention is made of conceptual support. In contrast, the poststudy example mentions the relevancy of the facts and indicates that referring to scientific laws can help to do this.

The overall mixture of results from this and the previous subsection suggests that the answer to inquiry question three question is “yes, but…”. In support of the affirmative part of that answer, the descriptive statistics cited above for responses to argument-related FR questions indicated student performance that was generally at a sufficiently high level relative to standards set by the College Board. Additionally, evidence from pre/post assessment comparisons suggests that the students did learn how to construct better arguments, but the lack of consistency and the importance of understanding unit-based content in argument construction may make these positive results *less* relevant to student effectiveness on the AP exam than hypothesized. This was particularly true when answering FR questions that demand a high level of conceptual understanding.

The connection between constructing a high-quality argument and sufficient understanding of discipline-based content may seem like a trivial point, but it has non-trivial pedagogical value. Effectively responding to argument-related AP Physics 1 exam questions requires a sufficiently understood concept-base that argument skill, by itself, simply cannot provide. Therefore, poor performance on such questions is not *necessarily* indicative a of lack of argument skill. Lesson design that does not take this insight into account runs the risk of frustrating students by asking them to do something they aren't fully prepared for.

5.0 Discussion

Findings associated with the action research described in this report support the idea that the development and exhibition of scientific argument skill is a complex issue. Although the answers to the inquiry questions were backed with evidence, I have yet to propose a conceptual justification that this evidence is linked to my claims. Here I will introduce a foundation that may provide that needed support. Such grounding should account for the generally positive but inconsistent nature of the movement of my students along two simultaneous learning progressions that were intermingled in this study. They are: (1) the development of argument skill apart from a domain-specific context and (2) the growth in conceptual understanding of domain-specific knowledge. Section 5.1 will discuss how the findings correlate with the expectations set by the review of literature, including new sources relating content-based conceptual understanding to argumentation, whose importance became apparent after the study began. Section 5.2 will introduce additional relevant literature-based ideas on problem representation, the inclusion of which was also prompted by the reflective work done during the study. I will first describe these ideas in section 5.2.1 and then apply them to the findings from this study in section 5.2.2. Section 5.3 summarizes the discussion while section 5.4 describes the limitations of the findings. The chapter closes with a short discussion on the implications of this study for my own practice in section 5.5 and suggestions for follow-up research in section 5.6.

5.1 Correlation of Findings with the Literature

The combination of intermingled argument-related and physics-related learning progressions that were a part of this study exemplifies an issue that argumentation research in education grapples with – that is, can pure argument skill be untangled from conceptual subject-matter understanding? The attempts identified earlier in this report (see sections 2.1.7 and 2.1.8) by some researchers to integrate assessment of subject-specific conceptual knowledge with evaluation of argument skill indicates that this issue is seen as important. However, it's also apparent that work still needs to be done to make the assessment of argument quality reliable within a complex content-specific argument construction process (Gotwals, Songer, & Bullard, 2012; Sampson & Clark, 2008). Although, investigators such as Osborne, Erduran, and Simon (2004, p. 1015) specifically refute that, "…students must acquire a knowledge of the major components of the scientific canon before they can engage in discourse activities that resemble or model those of the professional scientist", they also admit that, "…argument in a scientific context requires very specific knowledge of the phenomenon at hand and at least a feel for the criteria for evaluating scientific evidence. Without this resource, constructing arguments of quality will be severely restricted and hampered." The evidence identified in this study supports both insights.

Deana Kuhn (2010) also points to entanglement issues when she expresses concern that development of ambitious content knowledge simultaneously with argumentation skill development can be difficult for students. This was true for my students. However, she clearly asserts that the practice of this skill within a content-grounded context is important and should be done across diverse settings to be reinforced. A recurrent theme in her work is that intellectual skills stand separate from subject-specific contexts, but that context is relevant. Thus, Kuhn supports the intermingling of learning argument skills structurally with content-specific immersion experience in arguing (Kuhn 2010, Kuhn, 2012; Kuhn, Hemberger, & Khait, 2016). In this she differs from some who view argument skill development as not requiring a relevant context such as Larson, Britt, and Kurby (2009) *and* from those who view attempts at non-contextualized, direct instruction of argument, in a negative light such as Berland and Hammer (2012).

According to Kuhn (2010) and others (for example, Zohar & Nemet, 2002) both the practice of and the reflection on argumentation aids in the development of argument skill. The findings of this study support those two notions, but the general emphasis on practice, whether prompted by direct instruction or immersion, and reflection leave relatively unexplored the primary underlying role of subject-specific conceptual knowledge and its limiting effect on constructing high-quality arguments. Ogan-Bekiroglu and Eskin (2012, p. 1415) report that, "research focusing on the interplay between science understanding and argumentation practices is very rare." Does lack of subject-based conceptual understanding affect the expression of argument skill? The findings from my study (see sections 4.1, 4.2, and 4.3) prompt me to say yes. de Lima Tavares, Jimenez-Aleixandre, and Mortimer (2010) reported similar results on the connection between conceptual understanding of biological evolution and argumentation by 12th grade students in a Brazilian high school. Additionally, analysis by von Aufschnaiter and colleagues (2008) supports the claim that while construction of a high-quality argument with a low-level of knowledge is possible, it is extremely unlikely. That fairly obvious fact, they say, has pedagogical ramifications for instruction. That is, if instruction that includes argumentation is designed to promote domain-specific learning, a sufficient base of knowledge must be available to the student practitioner and that "…appropriate use of this pedagogy requires a more careful consideration of the interrelationship between the content and process of an argument" (von Aufschnaiter, et. al, 2008, p. 128).

My research, on an admittedly small sample of students, points to a connection between physics-specific conceptual understanding and expression of skill in argument construction that needs to be clarified. It may be true, in a non-science setting, that argument skill can be exhibited and evaluated in a decontextualized way. But arguments constructed in such a circumstance will not have value from a science-learning perspective other than as practice for later application. That assertion doesn't negate Kuhn's claim (2010, p. 822) that, "…the skills of argument are fundamental intellectual skills, worthy of attention in their own right". Instead, it prompts the more important question of how they stand relative to each other in a science class whose culture values argumentation. My findings indicate that domain-specific conceptual understanding provides more than a context for argument construction. It supplies an essential foundation. For scientific argument skill to be exhibited, a minimum level of appropriate domain-specific expertise must be present for the student to rely on. This is not to say that student knowledge of argument construction cannot develop as a separate domain of knowledge *within* the activities of a science course or that explicit instruction designed to foster that development should be avoided. Instead, the implication is that *meaningful* practice of such knowledge requires conceptual understanding from within another domain. For my purposes that domain is physics.

Perhaps it's instructive to clarify this point by way of a question. For argument-related activities in a physics course, is domain-specific knowledge in service of the development of argument skill or is argument skill in service of the learning of domain-specific knowledge? This question can be answered without denying the separate nature of each. In other words, they needn't be seen as having to be tangled together. However, claiming one option or the other doesn't allow for them to be seen as equal partners functioning in parallel. So, unless one declines to answer, or sees the question itself as misguided, the primacy of one above the other is implied in either response. The findings of my research generally support the conclusion that, in a physics course, argument skill is in the service of learning subject-specific content. While students were shown to improve their ability to construct arguments, the inconsistency of their performance indicates that developed argument skill, in general, does not guarantee the construction of high-quality physics-based arguments. Argument structure evaluations throughout this investigation provide ample evidence of this (see sections 4.1 to 4.3). I interpret the findings to indicate that unit-based subject-matter is both the atmosphere within which and the ground upon which argument construction activity occurred. Argument skill, exhibited within my course for the purpose of learning physics principles, was not meaningfully practiced apart from that context. Conversely, in courses such as AP Physics 1, subject-related conceptual understanding can be developed apart from the practice of argument skill in a wide variety of meaningful ways. This difference is a crucial insight that points to the primacy of concept development in a physics course. The same would be true in reverse if a course on argument construction utilized physics concepts in certain activities. In that case, domain-specific knowledge would be in the service of the development of argument skill.

5.2 Insights Using Problem Representation Concepts

A possible way to conceptualize the relationship between argument skill and the understanding of physics principles is supported through work done on problem representation by Schoenfeld and Herrmann (1982) and Chi and colleagues (1981). In the following two sub-sections, I will overview their insights and discuss how the findings of this investigation relate to them.

5.2.1 Problem Representation

Schoenfeld and Herrmann (1982) were interested in the differences between how novice learners and experts mentally represent problems in mathematics. One of their important points is that experts in a particular domain of knowledge recognize *deep structure* in problems and perceive them, in part, according to the concepts needed to solve them. Conversely, novices recognize *surface characteristics* and perceive problems based on elements that are given in the problem itself. Although Schoenfeld and Herrmann's work is the mathematics domain, they point to work by Chi and colleagues (1981) who note this same distinction is found in the perception of physics problems. For example, the surface structure of a physics problem that asks a student to find the speed of a ball on a ramp would consist of the objects mentioned (or sketched) and the words used in the text of the problem. Any mental representation is limited when surface structure is the extent of perception, making those associated with physics problems unlikely to help construct solutions. However, when deep structure is perceived, the mental representation of a problem includes such cognitive structures as domain-based concepts and connections to past problem-solving experience. In the ball and ramp problem the associated mental structures include the concept of energy conservation, principles of force analysis, and memories of similar problems experienced, and perhaps solved, in the past.

Schoenfeld and Herrmann (1982) are careful to say that the characteristics of "deep structure" for mathematics are different from those in physics. In mathematics, deep structure is associated with methods, but for physics it is associated with principles. They claim that because mathematics students are exposed to similar methods in different math classes, they generally start higher-level courses with some degree of expert understanding of problem-solving. That is not true of physics students, who generally do not have more than basic understanding of the principles of physics prior to taking a physics course and often have misconceptions about them. Even so, when discussing the results of their research, Schoenfeld and Herrmann (1982, p. 491) claim that for both math and physics students the “…surface structure is a primary criterion used by novices in determining problem relatedness. Moreover, it verifies directly that students' problem perceptions change as the students acquire problem-solving expertise. Not only their performance, but their perceptions, become more like experts.” If this idea applies to scientific argument construction, it may imply that for novice physics students who don’t have a developed understanding of a particular topic, the construction of a high-level argument is close to impossible. For some, the development may be partial, giving rise to better, but still not fully-constructed arguments. Additionally, since new topics are introduced in each unit of a physics course, students return to novice status over and over again until a certain level of expertise is developed within the newly introduced concepts.

Chi and colleagues (1981) follow this general framework when they theorize that physics students perceive and then cognitively represent problems as experts when they can activate a mental schema that is principle-oriented. This activation is the start of a mental cascade of connections, triggered by perceiving surface characteristics, but resulting in a complex cognitive representation that leads to a correct solution to the problem. Novice physics students cannot activate a principle-oriented schema that hasn’t yet been developed, so they must rely on surface characteristics in the problem to activate other problem-solving schema they have established through experience. However, those schemas are not as useful as ones that are principle-oriented. Chi and her co-authors go on to claim that, “…we presume that once the correct schema is activated, knowledge - both procedural and declarative - contained in the schema is used to further process the problem in a more or less top-down manner. The declarative knowledge contained in the schema generates potential problem configurations and conditions of applicability for procedures which are then tested with what is presented in the problem statement."

5.2.2 Applying Problem Representation Concepts to this Investigation

If argument construction requires mental representation similar to that needed for solving problems, then it's not surprising that students who only perceive the surface characteristics of an argument prompt, create low-level arguments. I believe the cognitive mechanism described above was on display during the pre/post cognitive appraisal interviews of this investigation when the students were asked to respond to claims given both data and a list of concepts related to those claims. The content of the mental schema possibly activated by the prompt was dependent upon several factors, the most important being the conceptual understanding of the primary scientific idea(s) associated with the claim. At the same time, activated schema content was possibly associated with conceptual understanding of the argument process and structure required to formulate an adequate response. If neither of these activations resulted in rich schema content, the student appeared confused and fumbled around until making an unsupported assertion. Here is an example of such a response to precourse interview question one, claim one, made by student 860:

> "Ok, um – Objects that are light in color – there would be object A, B, and E. So, they're all over 85 degrees to start. Um, object A heats up 40 degrees, C 18, and E is 29. That's a clear – much higher – Oh, wait – oops – ah, silver, yellow. OK, so it'd be A, D, and E all except for C heat up faster. A, D, and E and they come off with 85, 85, and 85. OK. They all start (ah) at about the same. If you watch it float in water - now. I would have to say I agree with that. C does not but the other A, and E are definitely, um, definitely change with that temperature that it goes up. OK. So, A, D, and E. So, density is all more. Density is

> higher based on the color. And the temperature – all above 85 degrees – I'm gonna say I agree with that."

If, however, the schema content was primarily associated with argument process and structure, with little or no conceptual content association, the response was likely to be an incomplete argument that contained some amount of factual evidence without justification. Fully constructed arguments required that schema content include sufficient information and guidance along both dimensions. Examples fitting each of these descriptions can be found in chapter four.

Additional data supporting the proposition that problem representation played a role in argument construction can be gathered from the quantitative structure analysis performed on the precourse and poststudy interviews. For example, when asked to respond to claims related to the concept of thermal conductivity on the precourse interview, the students scored consistently lower than on those related to density. This may indicate that students had a generally better understanding of the concept of density, an idea commonly discussed in science classes at many grade levels, than they had of thermal conductivity. Poststudy interviews showed the same relative results even though argument skill was shown to have been improved over the course of the study.

Further support for this idea comes from the analysis of student responses to the additional physics-related question on the poststudy interview (Q3) as average performance on Q3 was higher than that of the other two questions. The familiarity and conceptual comfort gained through exposure to mechanics ideas in AP Physics 1 possibly provided important schema content to the group as a whole. While the evaluation of this question cannot be compared for pre-to-post improvement, the average results are in line with the expectation that students make better arguments when they are based on solid conceptual understanding.

Finally, it was noted in the qualitative interview analysis that, on average, students who "used concepts" to a high degree for any particular interview response were more likely to be evaluated at a higher level. Students who relied of the use of facts generally had lower evaluations. I think it's fair to claim that the latter is more aligned with "surface characteristics" and the former with "deep structure" as described by Schoenfeld and Herrmann (1982). Students who perceive deep structure most likely express this perception with increased talk about concepts, while those who rely on surface characteristics do so with increased talk about facts.

It's reasonable to assume that the cognitive schema associated with argument process and structure were enriched by exposure to the argument-related instructional activities during this investigation. The data reported in chapter four support the conclusion that the course instruction on argumentation had an effect on argumentation skill as the calculated effect sizes, while not always high, pointed to some degree of group level improvement. Although not directly tested, it is also possible that the effectiveness of these strategies was enhanced by a study design that allowed for flexibility of instructional specifics. For example, at analysis point one, it was clear, based upon trends interpreted in student data, that many students needed to better understand the importance of conceptual justification in argument creation. Subsequently, live session activities and discussion forum prompts emphasized this idea. Additional areas, specifically addressed after other analysis points, were associated with sufficiency of evidence and conceptual accuracy. Perhaps the course instruction had a "leveling" effect on the strategies used on the whole as more students had access to similar appropriate cognitive schema and were more likely to construct better arguments. While it cannot be claimed that improvement occurred for each student in each argument construction scenario, group improvement was evident. A physics metaphor may help to clarify this idea. The average kinetic energy of an ideal gas is related to a very useful property called temperature, but it cannot be used to predict the behavior of any one individual particle in the gas at any particular moment. Likewise, the average group-level improvement noted in this study is a measure of the general argument construction abilities of the class but may not be indicative of the performance of any one individual student.

Many other findings reported here make sense in light of these insights. For example, even though there was small amount of overall group improvement from pretest to posttest, it didn’t occur for each student on each question. The nature of the pre/post test questions differed from all other assessments in this investigation in that there was an effort made to decontextualize them regarding subject-specific content. I think that question one achieved this to a higher (but still not perfect) degree than questions two and three (see Appendix A), but none were context-free. As was discussed earlier in this report, the cognitive resources needed to respond to questions two and three (formal-operational) were different from those used in question one (concrete-operational). Question one relied on descriptive hypothesis testing and was unlike that required by the other argument-related, highly contextualized, prompts used in this study. However, the causal hypothesis testing needed in questions two and three of the pre/posttests were much more similar to that required for the other investigation prompts and were also less decontextualized. So, although it was expected that it would be easier for students to construct arguments related to question one, the extent of decontextualization might have had the opposite effect as the students didn’t have concepts to rely on. Looked at from this perspective, it was not surprising that comparison of pre/posttest performance on questions two and three had medium effect sizes while question one did not. That finding may be the result of less decontextualization in those questions that allowed for activation of richer cognitive schema.

Findings related to the student-constructed artifacts are also supported by applying ideas associated with problem representation. General group-level argument structure improvement was found by unit, with medium and large effects sizes noted, relative to the pretest. If one interprets this finding through the lens of a learning progression associated with argument skill, then it's fair to say that students, as a group, did get better at argumentation. However, based upon the proposal that conceptual understanding grounds the process of argument construction in a science course, it's not surprising that argument skills are exhibited inconsistently when concepts are understood to various levels in different units. This was particularly apparent when considering the performance on FR argument construction where the demand for conceptual understanding was at its highest and changed from unit to unit. Although topics also changed by unit for lab-related arguments and FC rebuttals, those prompts were, in general, more highly focused than those presented in FR questions and were not influenced by the unique stressors associated with test-taking.

Given the picture painted thus far, one may expect that structure and accuracy evaluations should be correlated, as expression of accuracy is a logical measure of conceptual understanding. Generally, making accurate statements when constructing an argument implies that the cognitive schema activated by the prompt contains correct conceptual elements that aid in the process of argument creation. Without them, schema elements are limited to those associated with content-related surface structure, those related to argument construction, and perhaps others developed through related experience in problem-solving. As the conceptual understanding of argumentation develops within a student, a richer schema is produced that offers a higher level of service to whatever concept-related schema elements are available. This even applies to those that are incorrect, such as scientific misconceptions. Thus, students *can* make better arguments that include greater use of evidence without actually understanding why the evidence is linked to the claim. This dynamic is possibly responsible for some of the group-level improvement seen over time. More to the point, however, is that the relationship between structure and accuracy suggested here is supported by the strong positive associations found on the lab report data (Figure 8, p. 99), the FC data (Figure 9, p. 100) and the individual unit-average data (Figure 11, p. 101). While it's true that the same cannot be said of FR data (Figure 10, p. 100), the overall results imply something of practical pedagogical value that may be easily overlooked: Student performance when constructing a scientific argument relies on access to more than just content knowledge or argument skill alone.

5.3 Discussion Summary

The construction of a scientific argument requires conceptual understanding that is beyond that needed for typical problem-solving by physics students. It also demands conceptual understanding of what an argument is and how it should be constructed. For this investigation, a hybridized teaching strategy, using both direct instruction of and immersion in argumentation, designed to foster that understanding, was implemented. It's not surprising that, in general, the findings of this study did support the notion that my students improved their argument skill as a group based on argument structure evaluations. It's also not surprising, based on the assumption that both problem-solving and argument construction require adequate mental representation of a prompt, that they did so in a manner that was inconsistent unit by unit. What changed in each unit were the physics concepts required to construct the arguments, not the argument construction concepts. By this reasoning, student performance on the argument-related tasks in my AP Physics 1 course seemed to be dependent more on the level of conceptual understanding of the needed physics principle(s) than on the conceptual understanding of argument quality once it was achieved.

5.4 Limitations

The nature of this investigation suggests several reasons that its findings be considered with healthy skepticism. I acknowledge that this study's credibility can be questioned because students completed the work in remote settings, making the chance of outside help higher than ideal and that students may not have worked to their potential on all assessments. Additionally, although I attempted to be objective at every stage of the inquiry, I acknowledge that a desire to have my students succeed may have biased the evaluations. Using second coders to verify my evaluations helped to minimize this issue, but didn't eliminate it. Although a systematic approach to evaluation was used in order to promote analytical reflexivity, there was enough subjectivity in the methods that, even with well-designed rubrics, prejudicial results were possible. Additionally, the interpretation I assumed regarding the meaning of coverage percent data produced in the qualitative analysis of the cognitive appraisal interviews is debatable. I assumed that the percent coverage was a proxy for what was going on in the mind of a student when responding the interview prompts. Are there other ways to interpret this data? Undoubtedly.

Most importantly, the sample size ($N = 15$) was not large enough to make generalizable claims. For this reason, measures of statistical significance were not discussed in this report. Instead, effect size was determined when possible. Although the analysis made use of data measured through various methods, each of which produced findings that supported each other, and the results were generally in line with expectations set by the literature review, the results were meant to simply benefit my own practice. Action research of this type is not designed to inform other practitioners. Instead, it is meant to provide insight for the teacher-researcher as part of an attempt at pedagogical improvement. Can the results of this study be duplicated and can the findings be applied to other groups and circumstances? Those two criteria are important for traditional research. They find a different expression in action research. Action researchers invite others to "…learn from and perhaps adopt or adapt what you have done to their practices. This fulfills criteria to do with dynamic transformational potential, because other people can learn from you and can see new possibilities for their own research (McNiff & Whitehead, 2010, p. 16)." This summarizes my hope for this investigation beyond my own self-improvement and recognizes the limited reach of the findings.

5.5 Implications for Practice

As was mentioned at the beginning of this report, I have viewed each new school year as an opportunity for pedagogical improvement. The findings of this study support two adaptations of practice that may produce positive results in future iterations of my AP Physics courses. The first is associated with an increased confidence in the decision to actively promote a course culture that fosters the development of scientific argumentation. In other words, I have learned, in a more than an anecdotal way, that time taken up by argument-related instructional strategies is worth something. The data indicated that students do develop higher levels of argument skill when exposed to the instructional strategies outlined in this study. Because if this, I believe that the goals of the argument-related instruction set forth in this investigation - improved skill in disputative and deliberative argumentation - were achieved. It's reasonable to assume that in combination with effective instruction related to the learning of physics concepts, such strategies will benefit students when they take the AP Physics 1 exam and in other areas of academic life. It has been my experience that, too often, teachers of content-heavy science courses are hesitant to design lessons that aren’t specifically geared to subject-based learning. I include myself in this group. This hesitancy is somewhat understandable, as the pressure to cover the extensive set of learner objectives in a high-level science course is very real. It comes from students, parents, and administrators alike. There is an appeal to playing it safe, and not taking risks with successful traditional pedagogical approaches that are viewed as effective. However, most in-service teachers also know that improvement is always possible and comes by taking small pedagogical risks that may pay big dividends. Attempts to measure student progress through evaluation of various aspects of the “grasp of practice” (Ford and Foreman, 2006) are representative of this idea. I believe, based on the results of this investigation, that including both direct instruction and immersion strategies related to argumentation in AP Physics 1 aids in the internalization of that idea and can be reliably measured by argument structure evaluations within the context of the course. As such, the hybridized instructional strategy described in this report will continue to be utilized and refined to make grasping of authentic science practice more efficient and effective.

Secondly, for me, the results of this investigation have prompted deeper understanding of the complex relationship between scientific argumentation and content knowledge. Because of this, I will focus on transmitting this understanding to students in clear and helpful ways. For example, over many years I have emphasized the “big ideas” of physics with my students. I generally saw this as a way to promote long-lasting conceptual understanding of the fundamentals of physics that would benefit students well into the future. However, I now recognize the vital and more immediate nature of their internalization for argumentation specifically and problem-solving in general. As a result, the frequency with which I will engage my students in activities that promote this understanding will increase. Additionally, except for unit test assessment questions related to argument, I will utilize argument-related assignments only when a minimally sufficient level of concept understanding is assured. This will avoid student frustration with the assignment and increase the probability that exhibition of argument skill will promote the learning of course content.

5.6 Implications for Future Research

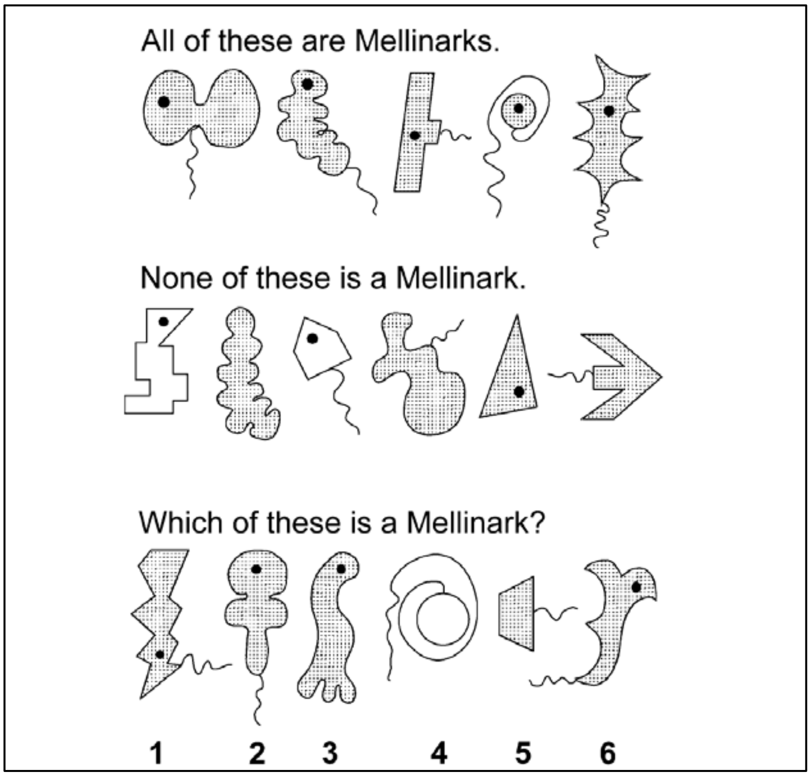
When designing this investigation, I was aided by the construction of a conjecture map (see section 3.3.3). On the map I identified two outcomes. The first was associated with the primary topic of this study: the construction of strong scientific arguments. The second predicted increased student awareness of the epistemic value of scientific argumentation. The results of this study support the view that inclusion of argument-related instructional strategies in AP Physics 1 benefits students when responding to argument-related questions. Do they also engender increased epistemic value for argumentation? I believe, like Kuhn (2001), that without valuing argumentation skill, it will not be utilized to the degree that it otherwise might. As a follow-up to the present investigation, systematically studying changes in the epistemic value of argumentation skill by students through self-reports would allow me to refine activities and make them more effective. Having students describe their experience when attempting to construct arguments in AP Physics 1 may provide insights into its relationship with course content and course-related motivation. A student generated meta-view of argumentation practice in my class may allow me to peek inside the heads of my students, gaining information that can be used to make the course freer of inhibiting factors that keep anxious yet qualified students from succeeding.

Additionally, it would be gratifying if other researchers, on a large scale, investigated the relationship between argument creation and the mental representation of an argument prompt. It was assumed in the previous discussion that the activation of cognitive schema in problem-solving was the same as that for argument construction except for the inclusion of schema elements from more than one domain of knowledge in the latter. Is that the case? If it is, how can awareness of that process benefit students? How can it be enhanced? If not, by what other cognitive mechanism(s) does the development of argument skill aid students in science-based argument construction?

Appendix A

Pre/Posttests

Pretest Question 1: Carefully examine the figures below and answer the question shown. In the space below the figures, fully explain your reasoning.

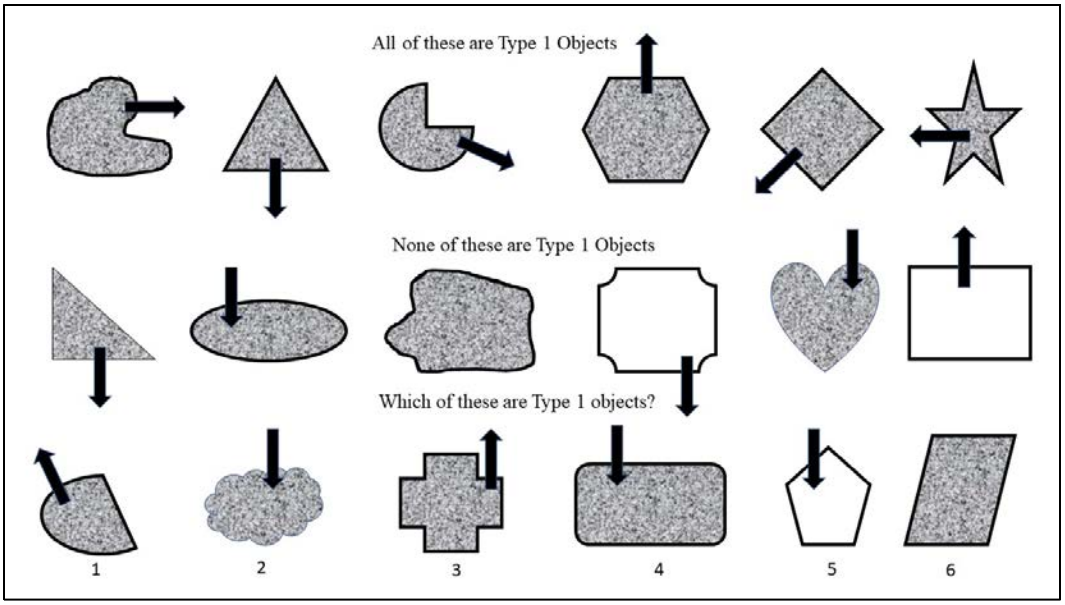


All of these are Mellinarks.

None of these is a Mellinark.

Which of these is a Mellinark?

Posttest Question 1: Carefully examine the figures below and answer the question shown. In the space below the figures, fully explain your reasoning.


All of these are Type 1 Objects
None of these are Type 1 Objects
Which of these are Type 1 objects?
1
2
3
4
5
6

Commentary on question 1 (*Category question*):

The “Melinark” question was identified by Lawson (2004) as a test for development into Piaget’s concrete operational stage. It was used in the Elementary Education Study (1974) cited by Lawson (2004). Using a category question as the first one was appropriate because it can indicate student development into the concrete operational stage. Student responses to this “category” question were evaluated using the rubric shown in Table 7 (p. 64). Sample responses at each level are shown in Table A1. Claim statements are bolded, evidence statements are italicized, and rationale (justification) statements are underlined.

Table A1. Sample responses for pretest question one

Argument Structure	Sample Response
Level 4	*By examining the objects it’s clear that all Mellinarks have an internal dot, a tail and are shaded. Non-Mellinarks are missing at least one of these features.* Having all three characteristics is essential for identifying an object as a Mellinark. **Objects 1, 2, and 6 are Mellinarks** because *they* show these characteristics, thereby meeting the standard for inclusion in the Mellinark category. *Objects 4 and 5 are missing at least one of these features.*

Level 3	**Objects 1, 2, and 6 are Mellinarks.** *They have internal dots, a tail, and are shaded. Non-Mellinarks are missing at least one of these characteristics.*
Level 2	**Objects 1, 2, and 6 are Mellinarks.** *They have an internal dot and a tail.*
Level 1	**Object 1, 2, and 6 are Mellinarks.**

Pretest Question 2: Imagine that you and a friend find a bag that is filled with many different objects. The objects are hidden in the bag until they are removed from it by your friend. Your friend reaches into the bag and pulls out seven objects one at a time. These objects are described below. Here is the information about the first seven objects removed from the bag:

Object	Shape	Color	Soft or Hard
1	Round Ball	Red	Soft
2	Square Block	Blue	Hard
3	Round Ball	Red	Hard
4	Square Block	Blue	Soft
5	Round Ball	Blue	Soft
6	Square Block	Red	Soft
7	Star	Blue	Hard

Your friend then removes another object from the bag (object 8). It is a red star. Do you think that object 8 is soft or hard? Fully explain your reasoning.

Posttest Question 2: You and several friends of yours like to collect items called “bots”. The group keeps a record of the characteristics of the bots on the pages of a notebook. There are many bots in the collection. Unfortunately, one of the entries (bot sample 36) on page five was partially unreadable because of a coffee stain. One of your friends says that she can determine the missing data based on the other entries shown on that page. Do you agree or disagree? Fully explain your choice.

Bot Sample	Class	Color	Texture
33	A	Gray	Metallic
34	B	Gray	Dull
35	B	Brown	Glassy
36	C	Gray	
37	C	Black	Metallic
38	C	Gray	Dull
39	A	Black	Metallic
40	A	Brown	Dull

Commentary on question 2 (*Predict the Characteristic of an Object*):

This question is scientific in nature but does not require discipline-specific science content knowledge to respond in a skilled way. Knowledge of mathematical probability may have helped the student in argument construction, but the evaluation of the quality of the argument did not depend upon it because only the structure of the argument was analyzed. This is the approach used by Osborne, Erduran, and Simon (2004) where only generic argument features were used to evaluate the level of argument skill exhibited by students. For a pretest, subject-matter content should play as small a role as possible.

The rubric shown in Table 7 (p. 64) was again used for the analysis of responses to question 2. Table A2 shows sample responses for each skill level. As above, claim statements are bolded, evidence statements are italicized, and rationale (justification) statements are underlined. Of course, many different responses could be evaluated at the same level. For example, another level four response could have been, “One cannot predict if object 8 is hard or soft. There is no correlation between shape, color, and hardness shown in the data table. In science, predictions may be able to made on the basis of observed patterns, but there are no patterns in the data from which to make a scientific prediction in this case."

Table A2. Sample responses for pretest question two

Argument Structure	Sample Response
Level 4	**Object 8 is soft** *because there are already more hard objects among the first seven* and it is likely that the number of objects with specific characteristics is evenly distributed if the bag was randomly filled.
Level 3	**Object 8 is soft** *because there are already more hard objects among the first seven.*
Level 2	**Object 8 is soft** *because stars are always soft.*
Level 1	**Object 8 is hard**

Questions similar to this are suggested by Lawson (1978) in his test of formal reasoning designed for high school- and college-aged students. His test consists of 15 items that require a student to choose the best answer from a list of options and subsequently write an explanation of why they chose that option. Each of the items are associated with Piaget's formal operational stage of cognitive development. According to Lawson (1978, p.12), "Formal operations include those reasoning processes that guide the search for and evaluation of evidence to support or reject hypothetical causal propositions. These operations are used in the isolation and control of variables, the combinatorial analysis of possible causal factors (combinatorial reasoning), the weighing of confirming and disconfirming cases (correlational reasoning), the recognition of the probabilistic nature of phenomena (probabilistic reasoning), and the eventual establishment of functional relationships between variables (proportional reasoning)."

Pretest Question 3: Read through the following information and respond to the prompts that follow:

Two groups of high school science students gathered data from an experiment that could be characterized by several different variables. Among these variables are the number ticks and tocks. The students could control the number of ticks and then measure the number of tocks. Data from five trials for each of the groups is shown here.

Group 1			Group 2	
Ticks	Tocks	Trial	Ticks	Tocks
7.0	14.3	1	3.1	6.2
10.0	20.0	2	6.8	13.6
16.5	33.1	3	11.2	22.4
23.7	45.9	4	18.9	37.8
37.4	75.0	5	22.8	45.6

Choose one claim from the following list:

A. The results from both groups are in agreement.

B. The results from the two groups don’t agree.

C. It’s unclear as to whether the results agree or disagree.

In the space below, ***support your choice as fully as you can***.

Posttest Question 3: Read through the following information about a hypothetical experiment and respond to the prompts that follow. Two groups of students performed experiments to determine if a relationship exists between two measurements taken on a specific physical system. They each conducted five trials that involved two measurements per trial. For each trial, the first measurement was controlled (Measurement A) and the second measurement was not controlled (Measurement B). The equipment used was not reported, but the data from the groups is shown below.

Group 1			Group 2	
Measurement A	Measurement B	Trial	Measurement A	Measurement B
7.0	21.1	1	3.1	9.3
10.0	29.8	2	6.8	20.4
16.5	49.4	3	11.2	33.6
23.7	71.3	4	18.9	56.7
37.4	112.1	5	22.8	68.4

Choose one claim from the following list:

A. The results from both groups are in agreement.

B. The results from the two groups don't agree.

C. It's unclear as to whether the results agree or disagree.

In the space below, ***support your choice as fully as you can***.

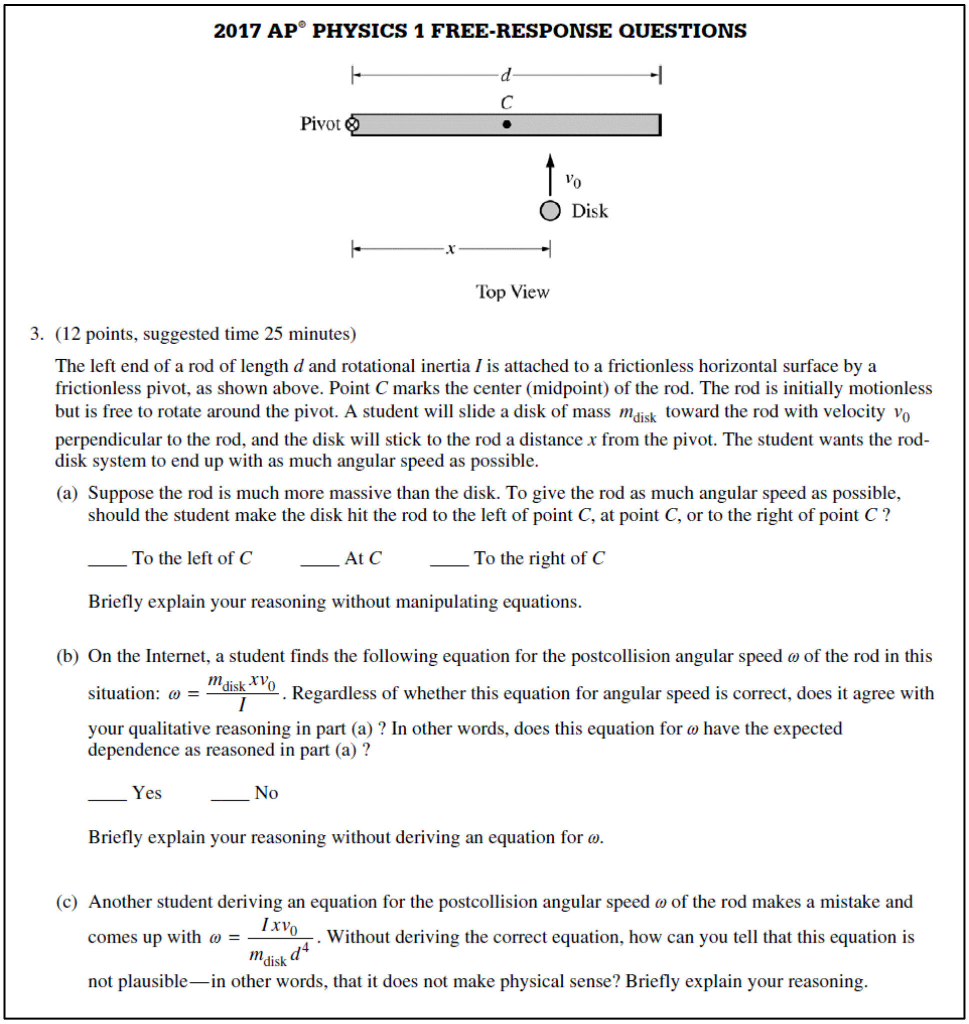
Commentary on question 3 (*Comparison*):

Like Question 2, this item is scientific in nature but did not require physics-specific subject matter knowledge to offer a valid response. It did require knowledge of mathematics, such as proportional reasoning. It is similar to one used by Luben, et.al., (2010, p. 2165) in a pretest designed to establish a baseline of argument skill exhibited by high school students prior to a group activity. Examples of pretest responses that would be evaluated at each level, using the structure rubric in Table 7 (p. 64), are shown in Table A3. Once again, claim statements are bolded, evidence statements are italicized, and rationale (justification) statements are underlined. Note that the claim *choice* is not evaluated. For example, another level four response choosing a different claim may have been, "**It's unclear as to whether the results agree or disagree** because not enough information is provided that details the methods used to make the measurements and the level of expertise of the students when making the measurements with given devices. *Although the data does show an exact proportional relation between ticks and tocks for group 2, the results from* *group 1 may show* that other factors (variables not measured) can affect the relationship between ticks and tocks or that group 2 did not make the measurements with care. In fact, they may have been biased by the expectation of an exact proportion."

Table A3. Sample responses for pretest question three

Argument Structure	Sample Response
Level 4	**The groups' data agree** *because the number of tocks is about double the number of ticks in each trial for each group*. The slight variation from this relationship shown in the data from group 1 can be due to many issues associated with making accurate measurements as is typical for science experiments.
Level 3	**The groups' data agree** *because the number of tocks is about double the number of ticks in each trial for each group.*
Level 2	**The group's data disagree** *because the values are not the same for each trial*.
Level 1	**The groups' data disagree.**

Question three also closely aligns with many of the free-response items used by the College Board on recent AP Physics 1 exams. The question shown in Figure A1, taken from the 2017 administration of the AP Physics 1 exam (College Board, 2017b), is typical in this regard. This example shows two often used short-answer question types. Parts (a) and (b) require that the student choose a claim about a physical situation and briefly explain the reasoning behind the choice. Part (c) presents a "false claim" that the student must refute through reasoning. This example differs from the ticks vs. tocks question because it requires physics experience that my students lacked at the time of the pretest administration. But the strong similarity of (1) choose a claim, followed by (2) explain your reasoning, is the same for the *ticks vs. tocks* question and many items on the AP Physics exam over the past few years.



2017 AP® PHYSICS 1 FREE-RESPONSE QUESTIONS

d
C
Pivot
v_0
Disk
x
Top View

3. (12 points, suggested time 25 minutes)

The left end of a rod of length d and rotational inertia I is attached to a frictionless horizontal surface by a frictionless pivot, as shown above. Point C marks the center (midpoint) of the rod. The rod is initially motionless but is free to rotate around the pivot. A student will slide a disk of mass m_{disk} toward the rod with velocity v_0 perpendicular to the rod, and the disk will stick to the rod a distance x from the pivot. The student wants the rod-disk system to end up with as much angular speed as possible.

(a) Suppose the rod is much more massive than the disk. To give the rod as much angular speed as possible, should the student make the disk hit the rod to the left of point C, at point C, or to the right of point C ?

____ To the left of C ____ At C ____ To the right of C

Briefly explain your reasoning without manipulating equations.

(b) On the Internet, a student finds the following equation for the postcollision angular speed ω of the rod in this situation: $\omega = \dfrac{m_{\text{disk}} x v_0}{I}$. Regardless of whether this equation for angular speed is correct, does it agree with your qualitative reasoning in part (a) ? In other words, does this equation for ω have the expected dependence as reasoned in part (a) ?

____ Yes ____ No

Briefly explain your reasoning without deriving an equation for ω.

(c) Another student deriving an equation for the postcollision angular speed ω of the rod makes a mistake and comes up with $\omega = \dfrac{I x v_0}{m_{\text{disk}} d^4}$. Without deriving the correct equation, how can you tell that this equation is not plausible—in other words, that it does not make physical sense? Briefly explain your reasoning.

Figure A1. Question 3 on the free-response section of the 2017 AP Physics 1 exam (College Board, 2017b)

Appendix B

Interview Protocols

Introductory Script:

Thanks so much for agreeing to be interviewed for my research project! I'm very excited to meet with you today. I want to learn from you how I can improve my teaching. Hopefully, this interview will help to create a partnership that will promote improved learning in the AP Physics 1 course. Your answers to the questions I ask today will allow me to understand how students think about certain types of questions. I anticipate that today's interview will take 20 to 30 minutes to complete.

Do you have any questions before we begin? [Field questions, or say I'll reach back after consulting with Ellen Ansell, my advisor.].

Interview Script:

Opening Script: I know that it might feel a bit awkward to work at solving a problem while we are interacting via a video conference – so I want to make you feel at ease when answering the questions during this interview. It's important for you to know I can learn from this experience regardless of how you answer the questions. Remember, that I'm interested in your thinking and that this interview will have no bearing on your grade in this course.

[Provide the interviewee with the information needed to respond to question 1 and read it aloud.] The objects described in this chart have all been sitting for several hours on a table in a room. The air in the room is kept at a constant 85 °F. All the objects are spherical (round). Objects

A, C, and E, all have the same mass. Objects B and D have twice the mass of object A. Object C is the least dense of the objects. All other objects have densities that are multiples of the density of object C. Recall the following concepts from your background in math and chemistry:

A. Density is the ratio between mass and volume of a substance.

B. Volume is the amount of “space” something takes up.

C. The rate at which an object changes temperature is controlled by the microscopic structure of the object.

D. You feel that an object is cold because your body is losing energy to that object.

E. The size of a sphere is based upon its radius.

F. The color of an object is based upon the light that is reflected from it.

G. Objects float or sink in water based on their density relative to water.

I know that there is a lot of data in the chart, so I am going to give you a few minutes to look it over. Let me know when you are ready to begin.

Object	Mass	Density	Color	Temperature (°F)	Does the object float in water?	How it feels when touched	Temperature change when placed in 150°F oven for 15 minutes
A	m	10d	Silver	85.2	No	Cold	+40
B	2m	2d	Brown	85.1	Yes	Warm	+9
C	m	d	Brown	84.9	Yes	Cool	+18
D	2m	4d	Yellow	85.1	Yes	Warm	+6
E	m	15d	White	85.0	No	Cold	+29

Question 1: Based on the information you are given here, consider each of the claims I will show you, one at a time. Your job is to decide if you agree or disagree with the claims and verbally

explain your choice as completely as you can. Take your time when responding. You may write notes if it helps you to decide. As you consider your choice, verbally express your thinking process as completely as you can. [Repeat the final sentence for emphasis, then show claim 1 and read it aloud.]

Claim 1: Objects that are light in color (such as the silver, yellow, or white objects) heat up faster than objects that are dark in color (such as the brown objects).

□ Agree □ Disagree

[When the interviewee has finished responding to claim 1, show claim 2 and read it aloud.]

Claim 2: Two of the objects are the same size (radius).

□ Agree □ Disagree

[When the interviewee has finished responding to claim 2, show claim 3 and read it aloud.]

Claim 3: Objects that feel cold are better conductors of heat.

□ Agree □ Disagree

[When the interviewee has finished responding to claim 3, show claim 4 and read it aloud.]

Claim 4: Objects with greater density conduct heat at a higher rate.

□ Agree □ Disagree

[Once the student has completed responding to claim 4, provide the information needed to respond to question 2 and read it aloud.] The properties for object F (also spherical) are now added to the chart except for the "float in water" and the density data. The size (radius) of object F is identical to that of object C.

Object	Mass	Density	Color	Temperature(°F)	Does the object float in water?	How it feels when touched	Temperature change when placed in 150°F oven for 15 minutes
A	m	10d	Silver	85.2	No	Cold	+40
B	2m	2d	Brown	85.1	Yes	Warm	+9
C	m	d	Brown	84.9	Yes	Cool	+18
D	2m	4d	Yellow	85.1	Yes	Warm	+6
E	m	15d	White	85.0	No	Cold	+29
F	4m		Silver	84.9		Cool	+15

Question 2: Does object F float in water? Write your argument on a sheet of paper. Be as complete as possible. Let me know when you are done.

□ Yes □ No

[When the student indicates that the work is complete, ask them to read the argument aloud.]

[The next section of the protocol, up to the final question, was used only in the poststudy interview. Provide the interviewee with the following information and read it aloud.] The objects described in this next chart are rectangular blocks of uniform density sliding on a straight, friction-free, track along a direction defined by the x-axis of a coordinate system. All of the objects have the same length (along the x-direction), but can differ in height and width. The position of the front edge of each block and its velocity data were measured at one specific instant in time. The acceleration data, however, were measured many times and found to be constant for each block during the entire experiment. Recall the following concepts from our course:

A. Newton’s second law of motion relates the net force on a given mass with its acceleration.

B. Kinetic energy is based upon mass and speed.

C. The momentum of an object is based on its mass and velocity.

D. Density is the ratio between the mass and the volume of a substance.

E. Velocity is the rate at which an object changes its position.

F. Acceleration is the rate at which an object changes its velocity.

Again, I know that there is a lot of data in the chart, so I am going to give you a few minutes to look it over. Let me know when you are ready for the respond to a claim.

Object	Mass	Position	Velocity	Acceleration	Kinetic Energy	Density
1	m	x	v	3a/2	K	d
2	2m	x/2	v/2	a	K/2	2d
3	m	x/10	v/4	3a/4	K/16	3d
4	2m	0	v/8	a/4	K/32	D
5	m	2x	2v	2a	4K	2d

Question 3: Based on the information you are given, consider each of the claims I will show you. Again, your job is to decide if you agree or disagree with the claims and verbally explain your choice as completely as you can. Take your time when responding. You may write notes if it helps you to decide. As before, verbally express your thinking process as completely as you can.

Claim 1: The same net force is being exerted on two of the objects in the chart.

□ Agree □ Disagree

Claim 2: For the objects shown in the chart, the bigger the volume the bigger the kinetic energy.

□ Agree □ Disagree

[This is the end of the extra section added to the poststudy protocol]

[The interviewee is then asked the final question]

Final Question: Describe what you think it means to make a strong scientific argument.

[When the interviewee has finished responding to the final question, read the conclusion script.]

Conclusion Script:

That's all the questions I have. Thanks for participating in this interview. Your input provides a valuable way for me to understand how to make the courses at Physics Prep as effective as possible. I appreciate your help!

Commentary: Tasks similar to the one described above were used by Sampson and Clark (2009, p. 476). A modified version of these tasks was used by Sampson and Blanchard (2012), this time as part of the protocol for a cognitive appraisal interview. They interviewed thirty high school science teachers to ascertain their ability to use scientific argumentation as well as their feelings about using argument as a teaching strategy. As part of the cognitive appraisal interviews, the participants were asked to assess three alternative statements that could be used as explanations of various physical processes such as the melting of ice and the changing phase of the moon. After choosing the explanation they felt was best, they were asked to support their choice verbally. The teachers were then asked to choose one of the physical processes and, using the data provided, construct a written scientific argument for the explanation they chose. The teacher was asked to reflect/comment on the thinking used to complete the task and then share with the interviewer what makes some arguments better or worse than others. The interviewers didn’t probe with planned follow-up questions as much as allowed the participants to appraise their own thinking.

Appendix C

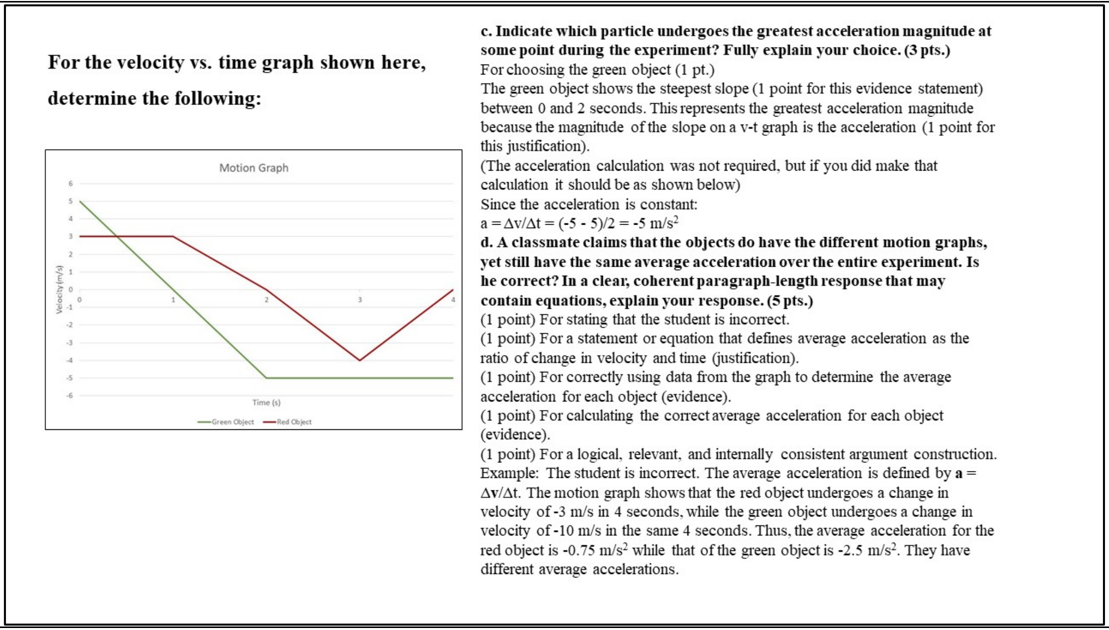
Prompts for Student-Constructed Argument Artifacts

<table>
<tr><th>Unit</th><th>False Claim Prompts</th></tr>
<tr><td>1</td><td>When an object is tossed vertically upward, its acceleration vector cannot be constant because the object returns to the Earth.</td></tr>
<tr><td>2</td><td>If I'm flying in an airplane (at constant speed) and I toss a ball straight upward it will obviously land behind me because I'm moving forward with the plane.</td></tr>
<tr><td>3</td><td>Friction always opposes motion, so the friction force can never cause something to accelerate if it is initially at rest.</td></tr>
<tr><td>4</td><td>When two equally strengthened individuals engage in a tug of war, and the rope does not move at all, each person still gets physically exhausted. That means that they each must do an equal amount of positive work on the rope.</td></tr>
<tr><td>5</td><td>A large spring is hung from the ceiling of a lab room. A box is then hung from the spring and you notice its periodic motion when released. The box is then removed from the spring. A second box (identical in size to the first) is then attached to the same spring. The periodic motion that you observe takes much longer to complete one cycle that with the first box. Your classmate claims that, "The second box must be have less average density than the first.</td></tr>
<tr><td></td><td>Lab Report Prompts</td></tr>
<tr><td>1</td><td>Examine the data shown below (Chart A) that could have been recorded for multiple attempts to make the measurements associated with one of the trials in this ramp experiment. This data may or may not be similar to the data you collected.
<table>
<tr><th>Attempt</th><th>Measured Time (s)</th><th>Measured Angle (degrees)</th></tr>
<tr><td>1</td><td>1.57</td><td>5.2</td></tr>
<tr><td>2</td><td>1.48</td><td>5.1</td></tr>
<tr><td>3</td><td>1.65</td><td>5.1</td></tr>
<tr><td>4</td><td>1.47</td><td>5.0</td></tr>
<tr><td>5</td><td>1.62</td><td>5.1</td></tr>
</table>
Chart A

Now read the two claims listed below. In the argument analysis section of the lab report identify which claim you think is correct based on the data in Chart A. Support your chosen claim with evidence and justify the connection between the claim and the evidence with text-based and mathematical reasoning. You need to utilize the concept of relative standard deviation, as discussed in the Error Analysis presentation, to satisfactorily support your claim.</td></tr>
</table>

	Claim 1: The uncertainty in the measurement of time is larger than the uncertainty in the measurement of angle. Claim 2: The uncertainty in the measurement of angle is larger than the uncertainty in the measurement of time.
2	The final step in this Forces and Motion Lab is to create one valid scientific argument based on what you learned in the lab. Recall, that a valid scientific argument contains a claim, evidence, and justification.
3	Finally, construct an argument (in paragraph form), that makes a claim about the concept of friction based upon what you learned in this Coefficient of Friction Lab. Be sure to include evidence from the lab and a conceptual justification.
4	Imagine that a fellow classmate told you that the he did an experiment involving two masses in a head on collision similar to the one you just did. One mass (10 kg) was at rest and then was struck by another mass (0.5 kg) such that they stuck together. He didn't tell you the incoming speed of the smaller mass, but he claims that the total kinetic energy of the system was the same after the collision as before. Create an argument (using evidence from the simulation) that refutes your friend's claim.
5	Design experiments to support or rebut the following claims: 1. The period of a pendulum is related to the square root of its length. 2. The period of a pendulum is independent of its mass. 3. The period of a pendulum depends upon the acceleration of gravity. 4. The pendulum is isochronous (period is independent of amplitude). Use the data you collect, and the theory that you know from the previous presentation, to construct a fully developed, paragraph-length, coherent scientific argument that supports or rebuts one of the above claims. Make sure that all components of your argument are related, with no extraneous data or concepts included. In other words, make sure that the evidence supports your claim AND that the concepts you discuss make it clear how the evidence is connected to the claim.
	Free-Response Prompts and Point-Based Rubrics

1



For the velocity vs. time graph shown here, determine the following:

Motion Graph
Velocity (m/s)
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
0
1
2
3
4
Time (s)
Green Object
Red Object

c. Indicate which particle undergoes the greatest acceleration magnitude at some point during the experiment? Fully explain your choice. (3 pts.)

For choosing the green object (1 pt.)

The green object shows the steepest slope (1 point for this evidence statement) between 0 and 2 seconds. This represents the greatest acceleration magnitude because the magnitude of the slope on a v-t graph is the acceleration (1 point for this justification).

(The acceleration calculation was not required, but if you did make that calculation it should be as shown below)

Since the acceleration is constant:

$a = \Delta v/\Delta t = (-5 - 5)/2 = -5$ m/s^2

d. A classmate claims that the objects do have the different motion graphs, yet still have the same average acceleration over the entire experiment. Is he correct? In a clear, coherent paragraph-length response that may contain equations, explain your response. (5 pts.)

(1 point) For stating that the student is incorrect.

(1 point) For a statement or equation that defines average acceleration as the ratio of change in velocity and time (justification).

(1 point) For correctly using data from the graph to determine the average acceleration for each object (evidence).

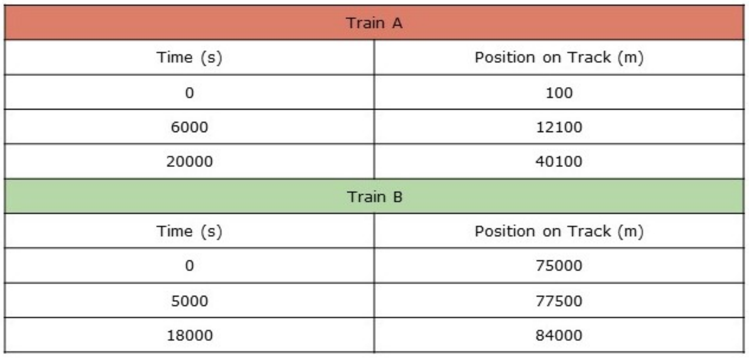
(1 point) For calculating the correct average acceleration for each object (evidence).

(1 point) For a logical, relevant, and internally consistent argument construction.

Example: The student is incorrect. The average acceleration is defined by $\mathbf{a} = \Delta\mathbf{v}/\Delta t$. The motion graph shows that the red object undergoes a change in velocity of -3 m/s in 4 seconds, while the green object undergoes a change in velocity of -10 m/s in the same 4 seconds. Thus, the average acceleration for the red object is -0.75 m/s^2 while that of the green object is -2.5 m/s^2. They have different average accelerations.

2

Train A and train B are heading in the same direction on a straight track. The chart below shows their front positions on the track at various times. Assume that the trains are not accelerating. Use the FOR wherein positive is to the right of the zero point (the train station).



Train A	
Time (s)	Position on Track (m)
0	100
6000	12100
20000	40100

Train B	
Time (s)	Position on Track (m)
0	75000
5000	77500
18000	84000

Will the trains collide within 10 hours? In a clear, coherent, paragraph-length response that may contain equations and mathematics, fully explain your answer.

(1 point) The trains will not collide within 10 hours (claim).

(1 point) The student states that train A is 74900 m behind train B at the start of the experiment **OR** states some equivalent use of data (evidence).

(1 point) The student states the definition of average speed as distance divided by time or $v = \Delta x/\Delta t$ **OR** the student creates symbolic kinematic equations that relate position and time (justification).

(1 point) Use the average speed equation to find that Train A is traveling at 2.0 m/s and that Train B is traveling at 0.5 m/s (relative to the Earth) **OR** the students states that Train A is traveling faster than train B (in the same direction) by 1.5 m/s **OR** the student shows a relative speed calculation such as $\mathbf{v}_{AB} = \mathbf{v}_{AE} + \mathbf{v}_{EB} = 2.0 – 0.5 = 1.5$ m/s **OR** the student uses kinematic relationships to set up an equality that allows for the determination of the collision time (evidence).

(1 point) The student calculates the time to collide: Δt to collide = (distance between the trains at the start)/relative speed. Using this equation, the time to collide is Δt = 74900/1.5 = 49935 second = 13.9 hours **OR** the student uses kinematics to determine the time to collide OR shows that train B is still ahead of Train A at t = 10 hours (evidence).

(1 point) For a logical, relevant, and internally consistent argument construction.

3

This question is part (b) of a three-part question on satellites. Part (a) asked the student to derive the equation that relates speed to orbital radius ($v = (GM/r)^{½}$) by using Newton's Second Law (**F** = m**a**).

	The following question refers to a satellite in circular orbit around the Earth. Assume no air resistance. The Earth has a mass of $5.98\text{x}10^{24}$ kg and a radius of $6.36\text{x}10^{6}$ m. Assume that the "center to center" distance for this orbit is $2.1\text{x}10^{7}$ m, and explain (in paragraph form that should include mathematics) why you think the satellite is or is not geosynchronous. (1 point) The period of the Earth's rotation is T = 86400 s (evidence). (1 point) The period of the satellite is determined by its speed and its orbital radius **OR** $T = 2\pi r/v$ OR an equivalent equation such as $v = 2\pi r/T$ (justification). (1 point) The speed of a satellite is related to its mass and its orbital radius **OR** $v = (GM/r)^{\frac{1}{2}}$ **OR** use of the correct use of an answer from a previous part of the question even if the work is incorrect (justification). (1 point) Use of the two equations described above. (justification) $2\pi r/T = (GM/r)^{\frac{1}{2}}$ $T = (2\pi r^{3/2})/(GM^{)\frac{1}{2}}$ $T = (2\pi(2.1\text{x}107)^{3/2})/[G(5.98\text{x}1024)^{]\frac{1}{2}}$ NOTE: If the equation $T^2 = 4\pi 2r^3/(GM)$ is stated and used correctly, the student earns the 3 previous points as it combines the previous three steps. (1 point) The period of the satellite is 30260 s **OR** an answer consistent with an incorrect equation from part (a) (evidence). (1 point) A geosynchronous satellite must have the same period as that of the Earth's rotation (justification). (1 point) The satellite is not geosynchronous because the period of the Earth's rotation is not the same as the orbital period of the satellite (claim).
4	Two equal mass rubber balls are made to collide with each other in several different experiments conducted by various groups of students who claim results as shown below. Use the FOR wherein positive is to the right. Experiment 1: Mass 1: A rubber ball (mass = m) with an initial speed of +1 m/s Mass 2: A rubber ball (mass = m) initially at rest Outcome of the experiment: The masses move in the same direction at different speeds (mass 1 has a speed of +0.3 m/s while mass 2 has a speed of +0.7 m/s). Experiment 2: Mass 1: A rubber ball (mass = m) with an initial speed of +1 m/s Mass 2: A rubber ball (mass = m) initially at rest Outcome of the experiment: The masses move in same direction at equal speeds of +0.5 m/s. Experiment 3: Mass 1: A rubber ball (mass = m) with an initial speed of +1 m/s Mass 2: A rubber ball (mass = m) initially at rest Outcome of the experiment: One mass remains at rest while the other moves at +1 m/s. Choose which of the following claims is true and construct a paragraph-length, coherent scientific argument that thoroughly explains your selection. Claim 1: One of the experiments is physically possible in the everyday world. Claim 2: None of the experiments are physically possible in the everyday world. (1 point) for indicating that claim 1 is true (claim). (1 point) for indicating that $p = mv$ and that $K = \frac{1}{2}mv^2$ (justification). (1 point) for indicating that momentum must be conserved during all collisions (justification). (1 point) for indicating that kinetic energy cannot be conserved in a real-life collision. Some kinetic energy must be lost (justification). (1 point) for indicating that experiment 1 is possible because momentum is conserved and kinetic energy is lost (evidence).

	(1 point) for indicating that experiment 2 is theoretically possible because momentum is conserved and kinetic energy is lost, but not physically possible because the results indicate a completely inelastic type of collision that requires an attachment mechanism that is missing on rubber balls (evidence). (1 point) for indicating that experiment 3 is not possible because although momentum is conserved kinetic energy is not lost (evidence).
5	A nearly massless, horizontal strut (L = 2.0 m) exists in deep space. The strut can be visualized as a thin pipe that is oriented along the x-axis of a coordinate system with its left end at the origin (x = 0). Three forces are applied to the strut and are measured using a force probe: Force #1: 7N acting at 20° at 0.7 m from the left end Force #2: 3N acting at 90° at 1.2 m from the left end Force #3: 5N acting at 90° at 1.5 m from the left end The angles are measured counterclockwise from the positive x-axis. Use the rotational frame of reference wherein positive is counterclockwise. Choose one of the claims shown below and construct a paragraph-length, coherent scientific argument that thoroughly explains your claim selection and the series of steps you made to support your choice. You can use both equations and text in your argument. Claim 1: The equilibrant force must be exerted to the right of the 5N force. Claim 2: The equilibrant force must be exerted to the left of the 5N force. (1 pt.) For indicating that the total torque must be zero for the system to be in complete equilibrium (justification). (1 pt.) For indicating that the total force must be zero for the system to be in complete equilibrium (justification). (1 pt.) For reporting the magnitude of the equilibrium force or reporting the magnitude of its y-component as needed in the determination of the equilibrium force position (evidence). (1 pt.) For indicating that the torque magnitude is determined by τ = Fy(lever arm) or τ = rFsinθ (justification). (1 pt.) For reporting the magnitude of the equilibrium torque (evidence). (1 pt.) For utilizing the torque equation to show that the position of the equilibrium force is 1.23 m from the left end. This point can still be earned by correctly using any error carried forward in the argument (justification). (1 pt.) For choosing claim 2 (or either claim consistent with the equilibrium position calculation made in the argument) (claim).

Appendix D

Live Session and Discussion Forum Activities Related to Argumentation

Unit	Direct Instruction Activity	Description
1	Live Session #1	Students were introduced to the structural components of scientific argumentation.
1	Recorded Lecture: Evidence-Based Scientific Argumentation	An in-depth exposition of the nature of scientific argumentation and how it differs from common everyday argumentation
1	Live Session #2	Discussion of expectation for lab report arguments and false claim rebuttals.
2	Live Session #3	Discussion of the importance of conceptual justification on a free-response problem.
2	Live Session #4	Discussion of appropriate and sufficient evidence when responding to the false claim from unit 1 and the importance of conceptual justification as making a link between the claim and the evidence.
3	Live Session #5	Dissection of a student-constructed argument post from the discussion forum activity in unit 2.
3	Live Session #6	Discussion of requirements for the lab-related argument in the coefficient of friction lab. Introduction of the discussion forum activity for unit 3 along with an example wherein conceptual, not logical, justification was shown to strengthen the argument.
5	Live Session #10	Teacher review of a discussion forum argument activity post from unit 4 with emphasis on argument coherency. Additionally, students were introduced to the unit 5 discussion forum argument activity with an example.
5	Live Session #12	Teacher review of argument-construction on free-response questions with emphasis on conceptual accuracy and its relation to creating a strong argument. Examples were given and discussed.

Unit	Immersion Activity	Description

1	Live Session #1	Students were asked to critique an argument both verbally and via chat comments.
1	Discussion Forum Activity on Reaction Time Activity	Students were asked to make posts that support, rebut, or extend of one of several teacher-made claims about a lab activity the students performed.
1	Practice Problems	Six scenario-based problems gave students a chance to construct and critique scientific arguments.
2	Live Session #4	Student were asked to critique an argument both verbally and via chat comments in light of the discussion of the false claim work from unit 1.
2	Discussion Forum Activity on Newton’s Laws Lab Activity	Students were asked to make posts that support, rebut, or extend of one of several teacher-made claims about a lab activity the students performed.
3	Live Session #5	Students were asked to identify areas of strength and weakness in a given argument in light of the discussion of the forum posts from unit 2.
3	Discussion Forum Activity on Inertial and Gravitational Mass	Students were asked to choose an everyday activity that includes the use of an object whose motion is observable (a football, a gymnast, a runner, etc.). They were then asked to construct an argument about how the object would behave differently if the inertial mass of the object had twice the magnitude of its gravitational mass. This assignment was designed to address a noted lack of conceptual justification in previously constructed artifacts.
4	Live Session #7	Students were asked to identify strengths and weaknesses in a selected argument post and follow-up response from the unit 3 discussion forum activity. Additionally, students were asked to identify evidence and rationale that could be used to debunk a pseudo-scientific experiment captured on video.
4	Live Session #8	Students were asked to collectively evaluate selected student-created argument artifacts from unit 3.
4	Discussion Forum Activity on Energy and Systems	Students were asked to post a claim to the discussion forum about a physical activity supplemented with the following: A list of external forces that act on your body during that activity, a flow statement of energy transformations that occur as a result of that activity, and a list of concepts that would be required to be understood if you were asked to construct a scientific argument about your claim.
5	Live Session #9	Students were asked to discuss and identify components in a selected student-constructed argument that was posted to the discussion forum in unit 4. This activity addressed a noted deficiency in responses to initial posts wherein evidence was often insufficient.

5	Discussion Forum Activity on Center of Gravity	Students were asked to imagine that they worked as a mechanical engineer for a consulting firm. Clients approach their company for advice about construction projects. The assignment was to describe a problem whose solution involves all, some, or one of the concepts of rotation, torque, and center of gravity and then construct an argument that would outline an approach to solving the problem.
5	Live Session #11	Students were asked to discuss a student-constructed artifact from the unit 5 discussion forum activity with a focus on conceptual accuracy (an area of concern noted on free-response argument constructions)

Appendix E

Chart of Structure and Accuracy Rubric Combinations

Coding Combination	Structure Components
	Bold = Claim
	Italics = Evidence
	Underlined = Justification
		Accuracy Components
		Bold = Accurate
		Italics = Inaccurate

Structure Level 1 Accuracy Level 0	**The more massive an object, the further it will slide along a surface once released at a given speed.**	*The more massive an object, the further it will slide along a surface once released at a given speed.*
Structure Level 1 Accuracy Level 1	**The more massive an object, the more difficult it is to move it from rest and to keep it moving at a given speed.**	**The more massive an object, the more difficult it is to move it from rest** *and to keep it moving at a given speed.*
Structure Level 1 Accuracy Level 2	**The mass of an object does not determine how far it will slide before coming to rest from a certain speed.**	**The mass of an object does not determine how far it will slide before coming to rest from a certain speed.**
Structure Level 2 Accuracy Level 0	**The more massive an object, the more difficult it is to keep it moving at a given speed.** *When I tried to slide a refrigerator across a floor, it took less force to get it moving than to keep it moving.*	*The more massive an object, the more difficult it is to keep it moving at a given speed. When I tried to slide a refrigerator across a floor, it took less force to get it moving than to keep it moving.*
Structure Level 2 Accuracy Level 1	**The more massive an object, the more difficult it is to move it from rest and to keep it moving at a given speed.** *This is apparent to anyone who has tried to move a refrigerator as it is very difficult to keep it sliding along the floor.*	**The more massive an object, the more difficult it is to move it from rest** *and to keep it moving at a given speed. This is apparent to anyone who has tried to move a refrigerator as it is very difficult to keep it sliding along the floor.*
Structure Level 2 Accuracy Level 2	**The mass of an object does not determine how far it will slide before coming to rest from a certain speed.** *I conducted an experiment with two wooden blocks to show that this was true*.	**The mass of an object does not determine how far it will slide before coming to rest from a certain speed. I conducted an experiment with two wooden blocks to show that this was true.**
Structure Level 3 Accuracy Level 0	**The more massive an object, the more difficult it is to keep it moving at a given speed.** *When I tried to slide a refrigerator across a floor, it took less*	*The more massive an object, the more difficult it is to keep it moving at a given speed. When I tried to slide a refrigerator across a floor, it took less*

	force to get it moving than to keep it moving. One can also observe this when pushing a heavy box up a long ramp. At first, it’s easy to push the box, but after a while it becomes really difficult. One doesn’t notice the same increase in difficulty for a lightweight box.	*force to get it moving than to keep it moving. One can also observe this when pushing a heavy box up a long ramp. At first, it’s easy to push the box, but after a while it becomes really difficult. One doesn’t notice the same increase in difficulty for a lightweight box.*
Structure Level 3 Accuracy Level 1	**The more massive an object, the more difficult it is to move it from rest and to keep it moving at a given speed.** *When I tried to slide a refrigerator across a floor, it took less force to get it moving than to keep it moving. One can also observe this when pushing a heavy box up a long ramp. At first, it’s easy to push the box, but after a while it becomes really difficult. One doesn’t notice the same increase in difficulty for a lightweight box.*	**The more massive an object, the more difficult it is to move it from rest** *and to keep it moving at a given speed. When I tried to slide a refrigerator across a floor, it took less force to get it moving than to keep it moving. One can also observe this when pushing a heavy box up a long ramp. At first, it’s easy to push the box, but after a while it becomes really difficult. One doesn’t notice the same increase in difficulty for a lightweight box.*
Structure Level 3 Accuracy Level 2	**The mass of an object does not determine how far it will slide before coming to rest from a certain speed**. *I conducted an experiment that measured the time it took for a 1 kg block moving at 1 m/s on table surface. I repeated the experiment with a 2 kg block. The results were identical.*	**The mass of an object does not determine how far it will slide before coming to rest from a certain speed. I conducted an experiment that measured the time it took for a 1 kg block moving at 1 m/s on table surface. I repeated the experiment with a 2 kg block. The results were identical.**
Structure Level 4 Accuracy Level 0	**The more massive an object, the more difficult it is to keep it moving at a given speed.** *When I tried to slide a refrigerator across a floor, it took less force to get it moving than to keep it moving. One can also observe this when pushing a heavy box up a long ramp. At first, it’s easy to push the box, but after a while it becomes really difficult. One doesn’t notice the same increase in difficulty for a lightweight box.* These results aren’t surprising because they are supported by the common-sense idea nothing can move forever as that would violate the laws of thermodynamics.	*The more massive an object, the more difficult it is to keep it moving at a given speed. When I tried to slide a refrigerator across a floor, it took less force to get it moving than to keep it moving. One can also observe this when pushing a heavy box up a long ramp. At first, it’s easy to push the box, but after a while it becomes really difficult. One doesn’t notice the same increase in difficulty for a lightweight box. These results aren’t surprising because they are supported by the common-sense idea nothing can move forever as that would violate the laws of thermodynamics.*
Structure Level 4 Accuracy Level 1	**The more massive an object, the more difficult it is to move it from rest and to keep it moving at a given speed.** *When I tried to slide a refrigerator across a floor, it took less force to get it moving than to keep it moving. One can also observe this when pushing a heavy box up a long ramp. At first, it’s easy to*	**The more massive an object, the more difficult it is to move it from rest** *and to keep it moving at a given speed. When I tried to slide a refrigerator across a floor, it took less force to get it moving than to keep it moving. One can also observe this when pushing a heavy box up a long ramp. At first, it’s easy to push*

	push the box, but after a while it becomes really difficult. One doesn't notice the same increase in difficulty for a lightweight box. These results are supported by Newton's 1st law of motion (the law of inertia).	*the box, but after a while it becomes really difficult. One doesn't notice the same increase in difficulty for a lightweight box. These results are supported by Newton's 1st law of motion (the law of inertia).*
Structure Level 4 Accuracy Level 2	**The mass of an object does not determine how far it will slide before coming to rest from a certain speed.** *I conducted an experiment that measured the time it took for a 1 kg block moving at 1 m/s on table surface to come to rest. I repeated the experiment with a 2 kg block. The results were identical.* Force analysis using Newton's 2nd Law of Motion supports this finding. The mass cancels out in the calculation ($F_f = ma$) because the friction force is partly determined by the mass ($F_f = \mu mg$).	**The mass of an object does not determine how far it will slide before coming to rest from a certain speed. I conducted an experiment that measured the time it took for a 1 kg block moving at 1 m/s on table surface to come to rest. I repeated the experiment with a 2 kg block. The results were identical. Force analysis using Newton's 2nd Law of Motion supports this finding. The mass cancels out in the calculation ($F_f = ma$) because the friction force is partly determined by the mass ($F_f = \mu mg$).**

Bibliography

Arnold, S., (2014). Assessing student learning online: overcoming reliability issues. In: Sampson D., Ifenthaler D., Spector J., Isaias P. (eds) *Digital Systems for Open Access to Formal and Informal Learning*. Springer, Cham.

Asterhan, C. S., & Schwarz, B. B. (2016). Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist*, 51(2), 164-187.

Avgerinou, M., Gialamas, S., & Tsoukia, L., (2014). I2Flex: The Meeting Point of Web-Based Education and Innovative Leadership in a K–12 International School Setting. In: Sampson D., Ifenthaler D., Spector J., Isaias P. (eds) *Digital Systems for Open Access to Formal and Informal Learning*. Springer, Cham.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (C. Emerson & M. Holquist, Trans.). Austin: University of Texas 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, 797–817.

Benson, A. D. (2003). Assessing participant learning in online environments. *New Directions for Adult and Continuing Education*, 2003(100), 69-78. doi:10.1002/ace.120

Berland, L.K., & Hammer, D. (2012) Students' framings and their participation in scientific argumentation. In: Khine M. (eds) *Perspectives on Scientific Argumentation*. Springer, Dordrecht

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, 94, 765–793. doi:10.1002/sce.20402

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, 93(1), 26-55. doi:10.1002/sce.20286

Berland, L. K., & Reiser, B. J. (2011). Classroom communities' adaptations of the practice of scientific argumentation. *Science Education*, 95(2), 191-216. doi:10.1002/sce.20420

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *The Phi Delta Kappan*, 86(1), 8-21. doi:10.1177/003172170408600105

Brown, A. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *The Journal of the Learning Sciences*, 2(2), 141-178.

Cavagnetto, A. R. (2010). Argument to Foster Scientific Literacy: A Review of Argument Interventions in K-12 Science Contexts. *Review of Educational Research, 80*(3), 336-371. doi:10.3102/0034654310376953

Cavagnetto, A. (2018). Personal conversation during internship at WSU.

Cavagnetto, A., & Hand, B. (2012). The importance of embedding argument in science classrooms. In: Khine M. (eds) *Perspectives on Scientific Argumentation*. Springer, Dordrecht

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121-152. doi:10.1207/s15516709cog0502_2

Clark, D. B., Sampson, V., Weinberger, A., & Erkens, G. (2007). Analytic frameworks for assessing dialogic argumentation in online learning environments. *Educational Psychology Review*, 19(3), 343-374. doi:10.1007/s10648-007-9050-7

College Board (2012). *AP Physics 1 and AP Physics 2 Curriculum Framework 2014-2015*. Retrieved from www.collegeboard.org.

College Board (2015a). *Student performance q & a: 2015 AP physics 1 free-response questions* [PDF file]. Retrieved from https://secure-media.collegeboard.org/digitalServices /pdf/ap/ap15_physics1_student_performance_qa.pdf

College Board (2015b). *The Paragraph-Length Response in AP Physics 1 and 2*. Retrieved from https://apstudent.collegeboard.org/apcourse/ap-physics-1/exam-practice

College Board (2016a). *Student performance q & a: 2016 AP physics 1 free-response questions* [PDF file]. Retrieved from 1https://secure-media.collegeboard.org/digitalServices /pdf/ap/ap16_physics1_student_performance_qa.pdf

College Board (2016b). *AP Physics 1: Algebra-based 2016 free-response questions*. Retrieved from1https://securemedia.collegeboard.org/digitalServices/pdf/ap/ap16_frq_physics1.pdf

College Board (2016c). *AP Physics 1: Algebra-based scoring guidelines*. Retrieved from https://secure-media.collegeboard.org/ap/pdf/ap16-sg-physics-1.pdf

College Board (2017a). *Chief Reader Report on Student Responses: 2017 AP® Physics 1 Free-Response Questions* [PDF file]. Retrieved from https://securemedia.collegeboard.org /digitalServices/pdf/ap/ap17-chief-reader-report-physics-1.pdf

College Board (2017b). *AP Physics 1: Algebra-based 2017 free-response questions*. Retrieved from https://apcentral.collegeboard.org/pdf/ap-physics-1-frq-2017.pdf?course=ap-physics-1

College Board (2019). *AP Physics 1 the exam*. Retrieved from https://apcentral.collegeboard.org/ courses/ap-physics-1/exam?course=ap-physics-1

Collins, A. (1992) Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.) *New directions in educational technology*. Berlin: Springer-Verlag.

Committee on Programs for Advanced Study of Mathematics and Science in American High Schools, & Council, N. R. (2001). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools: Report of the content panel for physics*. Washington: National Academies Press.

Crippen, K. J., Archambault, L. M., & Kern, C. L. (2013). The nature of laboratory learning experiences in secondary science online. *Research in Science Education*, 43(3), 1029-1050. doi:10.1007/s11165-012-9301-6

Dawson, V. M., & Venville, G. (2010). Teaching Strategies for Developing Students' *Science Education*, *40*(2), 133-148. doi:10.1007/s11165-008-9104-y

Dede, C. (2004). If Design-Based Research Is the Answer, What Is the Question? A Commentary on Collins, Joseph, and Bielaczyc; diSessa and Cobb; And Fishman, Marx, Blumenthal, Krajcik, and Soloway in the JLS Special Issue on Design-Based Research. *The Journal of the Learning Sciences,* 13(1), 105-114.

de Lima Tavares, M., Jiménez-Aleixandre, M., & Mortimer, E. F. (2010). Articulation of conceptual knowledge and argumentation practices by high school students in evolution problems. *Science & Education,* 19(6-8), 573-598. doi:10.1007/s11191-009-9206-6

Denscombe, M. (2014). *The Good Research Guide: For Small-scale Research Projects*. Maidenhead, Berkshire: McGraw-Hill Education.

Drew, C. (2011, January 7). Rethinking advanced placement. *The New York Times*. Retrieved from http://www.nytimes.com.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education*, 32(1), 268-291. doi:10.3102/0091732X07309371

Edwards, A. and Talbot, R. (1999). *The Hard-pressed Researcher: A Research Handbook for the Caring Professions,* 2nd edn. Harlow: Pearson.

Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, 20(4), 399-483. doi:10.1207/S1532690XCI2004_1

Erduran, S., & Aleixandre, M. (2008). *Argumentation in science education: Perspectives from classroom-based research*. Springer, Dordrecht. doi:10.1007/978-1-4020-6670-2

Essays, UK. (2018, November). *Criticism of Action Research*. Retrieved from https://www.ukessays.com/essays/education/research-methods-and-critique-action-research-in-education-education-essay.php?vref=1

Ford M. J., & Forman, E. A. (2006). Redefining Disciplinary Learning in Classroom Contexts. *Review of Research in Education, 30,* 1-32.

Fullerton, D. (2017). *What is the Difference Between AP Physics B and AP Physics 1 and 2?* Retrieved from https://www.educator.com/studyguide/physics/what-is-the-difference-between-ap-physics-b-and-ap-physics-1-and-2/

Garrison, D. R., & Cleveland-Innes, M. (2005). Facilitating cognitive presence in online learning: Interaction is not enough. *American Journal of Distance Education*, 19(3), 133-148. doi:10.1207/s15389286ajde1903_2

Genesee, F. (1985). Second language learning through immersion: A review ot U.S. programs. *Review of Educational Research*, 55, 541-561.

Gorard, S., Roberts, K., & Taylor, C. (2004). What kind of creature is a design experiment? *British Educational Research Journal*, 30(4), 577-590. doi:10.1080/0141192042000237248

Gotwals, A. W., Songer, N. B., & Bullard, L. (2012). Assessing students’ progressing abilities to construct scientific explanations. *Learning progressions in science*, 183-210.

Hanna, L. G. (2012). Homeschooling education: Longitudinal study of methods, materials, and curricula. *Education and Urban Society*, 44(5), 609-631. 10.1177/0013124511404886

Hatch, J. A. (2002). *Doing qualitative research in educational settings*. Albany, NY: SUNY Press.

Hickey, D. T. (2015). A situative response to the conundrum of formative assessment. *Assessment in Education: Principles, Policy & Practice*, 22(2), 202-223. doi:10.1080/0969594X.2015.1015404

Huang R., Yu L., Yang J. (2014). The Evolution of University Open Courses in Transforming Learning: Experiences from Mainland China. In: Sampson D., Ifenthaler D., Spector J., Isaias P. (eds) *Digital Systems for Open Access to Formal and Informal Learning*. Springer, Cham.

Iordanou, K. and Constantinou, C. P. (2015), Supporting Use of Evidence in Argumentation Through Practice in Argumentation and Reflection in the Context of SOCRATES Learning Environment. *Sci. Ed*., 99: 282–311. doi:10.1002/sce.21152

Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "doing the lesson" or "doing science": Argument in high school genetics. *Science Education*, 84(6), 757.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., & Bonney, C. R. et al. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, 42(3), 139-151.

Kelly, A. & Lesh, R. (2002). Understanding and explicating the design experiment methodology, *Building Research Capacity*, 3, 1-3

Kuang, C. (2011). *The 6 pillars of Steve Jobs's design philosophy*. Retrieved from https://www.fastcodesign.com/1665375/the-6-pillars-of-steve-jobss-design-philosophy.

Kuhn, D. (1991). *The skills of argument*. Cambridge, England: Cambridge University Press.

Kuhn, D. (2001). How do people know? *Psychological Science*, 12(1), 1-8. doi:10.1111/1467-9280.00302

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education*, 94(5), 810-824. doi:10.1002/sce.20395

Kuhn, D. (2012). Forward. In: Khine M. (eds) *Perspectives on Scientific Argumentation*. Springer, Dordrecht

Kuhn, D., Hemberger, L., & Khait, V. (2016). *Argue with me: Argument as a path to developing students' thinking and writing* (Second ed.). New York, NY: Routledge.

Kuhn, D., & Shaughnessy, M. E. (2004). An interview with Deanna Kuhn. *Educational Psychology Review*, 16(3), 267-282. doi:10.1023/B:EDPR.0000034197.75510.ef

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development*, 74(5), 1245-1260. doi:10.1111/1467-8624.00605

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*, 2nd enl. ed. University of Chicago Press.

Kunzman, R. (2012). Education, schooling, and children's rights: The complexity of homeschooling. *Educational Theory*, 62(1), 75-89. 10.1111/j.1741-5446.2011.00436.x

Larson, A. A., Britt, M. A., & Kurby, C. A. (2009). Improving students' evaluation of informal arguments. *The Journal of Experimental Education*, 77(4), 339-365. Retrieved from http://pitt.idm.oclc.org/login?url=https://search-proquest-com.pitt.idm.oclc.org/docview/217682819?accountid=14709

Lawson, A. E. (2004). The nature and development of scientific reasoning: A synthetic view. *International Journal of Science and Mathematics Education*, 2(3), 307-338. doi:10.1007/s10763-004-3224-2

Magrogan, S. (2014). Past, present, and future of AP chemistry: A brief history of course and exam alignment efforts. *Journal of Chemical Education*, 91(9), 1357.

Manz, E. (2015). Representing student argumentation as functionally emergent from scientific activity. *Review of Educational Research*, 85(4), 553-590. doi:10.3102/0034654314558490

Martinez, M. E., & Peters Burton, E. E. (2011). Cognitive affordances of the cyberinfrastructure for science and math learning. *Educational Media International*, 48(1), 17–26.

McDonald, C. V., & McRobbie, C. J. (2010). Utilising argumentation to teach nature of science. In B. J. Fraser, K. G. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education.* Dordrecht, The Netherlands: Springer.

McNeill, K. L. (2009). Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena. *Science Education*, 93(2), 233-268. doi:10.1002/sce.20294

McNeill, K. L., & Krajcik, J. (2007). Middle school students’ use of appropriate and inappropriate evidence in writing scientific explanations. In M. C. Lovett & P. Shah (Eds.), *Thinking with data: The Proceedings of the 33rd Carnegie Symposium on Cognition* (pp. 233 – 265). Mahwah, NJ: Erlbaum.

McNeill, K. L., & Krajcik, J. (2008). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching*, 45(1), 53-78. doi:10.1002/tea.20201

Meyer, K. A. (2014). *Student engagement online: What works and why*. Hoboken, New Jersey: John Wiley & Sons.

McNiff, J., & Whitehead, J. (2010) *You and Your Action Research Project*. (4th ed.). London; New York; Routledge Falmer.

National Research Council. (2002) *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools*. Washington, DC: The National Academies Press.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Committee on a Conceptual Framework for New

K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education*, 21(5), 553-576. doi:10.1080/095006999290570

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* Washington, D.C: National Academies Press.

Ogan-Bekiroglu, F., & Eskin, H. (2012). Examination of the relationship between engagement in scientific argumentation and conceptual knowledge. *International Journal of Science and Mathematics Education*, 10(6), 1415-1443. doi:10.1007/s10763-012-9346-z

Osborne, J. (2009). An argument for arguments in science classes. *Phi Delta Kappan*, 91(4), 62+.

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. doi:10.1002/tea.20035

Osborne, J., MacPherson, A., Patterson, A., & Szu, E. (2012). Introduction In: Khine M. (eds) *Perspectives on Scientific Argumentation*. Springer, Dordrecht

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, 340(6130), 320-323. doi:10.1126/science.1232065

Physics Prep. (2019). Retrieved from https://www.physics-prep.com.

Perkins, D., Jay, E., & Tishman, S. (1993). Beyond Abilities: A Dispositional Theory of Thinking. *Merrill-Palmer Quarterly*, 39(1), 1-21.

Pifarré M., Wegerif R., Guiral A., del Barrio M. (2014) Developing Technological and Pedagogical Affordances to Support the Collaborative Process of Inquiry-Based Science Education. In: Sampson D., Ifenthaler D., Spector J., Isaias P. (eds) *Digital Systems for Open Access to Formal and Informal Learning*. Springer, Cham

Rapanta, C., Garcia-Mila, M., & Gilabert, S. (2013). What is meant by argumentative competence? An integrative review of methods of analysis and assessment in education. *Review of Educational Research*, 83, 483–520. doi:10.3102/0034654313487606

Rehg, W. (2009). *Cogent science in context: The science wars, argumentation theory, and Habermas*. Cambridge, Mass: MIT Press.

Reiser, B., Berland, L, & Kenyon, L. (2012). Engaging students in the scientific practices of explanation and argumentation: understanding A Framework for K-12 Science Education. *The Science Teacher*, 79(4), 34+.

Rowe, E., & Asbell-Clarke, J. (2008). Learning science online: what matters for science teachers? *Journal of Interactive Online Learning*, 7(2), 75-104.

Ryu, S., & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education*, 96, 488-526.

Sandoval, W. A. (2014). Conjecture mapping: An approach to systematic educational design research. *The Journal of the Learning Sciences*, 23, 18-36.

Sampson, V., & Blanchard, M. R. (2012). Science teachers and scientific argumentation: Trends in views and practice. *Journal of Research in Science Teaching*, 49(9), 1122-1148. doi:10.1002/tea.21037

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*(3), 447-472. doi:10.1002/sce.20276

Sampson, V., Grooms, J., & Walker, J. P. (2011). Argument-Driven inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education*, 95(2), 217-257. doi:10.1002/sce.20421

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education*, 89(4), 634-656. doi:10.1002/sce.20065

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

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, 88(3), 345-372. doi:10.1002/sce.10130

Schoenfeld, A. H. (2006). Design experiments. In P. B. Elmore, G. Camilli & J. Green (Eds.), *Complementary methods for research in education*. Washington, DC: American Educational Research Association.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5), 484-494. doi:10.1037/0278-7393.8.5.484

Sher, A. (2009). Assessing the relationship of student-instructor and student-student interaction to student learning and satisfaction in Web-based Online Learning Environment. *Journal of Interactive Online Learning*, 8(2), 102-120.

Silverman, S. K. (2010, May). *Cognitive Appraisal Interviews for Surveys Embedded in Mixed-Methods Research*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Simon, S., Richardson, K., & Amos, R. (2012) The Design and Enactment of Argumentation Activities. In: Khine M. (eds) *Perspectives on Scientific Argumentation*. Springer, Dordrecht

Sloane, F. & Gorard, S. (2003) Exploring modeling aspects of design experiments, *Educational Researcher*, 31(1), 29-31.

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

von Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching,* 45(1), 101-131. doi:10.1002/tea.20213

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard, MA: Harvard University Press.

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

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.