EXPLORING THE RELATIONSHIP BETWEEN STUDENTS' LEVEL OF CONTENT KNOWLEDGE AND THEIR ABILITY TO ENGAGE IN SCIENTIFIC ARGUMENTATION USING STRUCTURAL EQUATION MODELING

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

ERIC WULFF

Dr. Patricia Friedrichsen & Dr. Troy Sadler , Co-Dissertation Supervisors

May 2019

The undersigned, appointed by the dean of the Graduate School,

have examined the dissertation entitled

EXPLORING THE RELATIONSHIP BETWEEN STUDENTS' LEVEL OF CONTENT KNOWLEDGE AND THEIR ABILITY TO ENGAGE IN SCIENTIFIC ARGUMENTATION USING STRUCTURAL EQUATION MODELING

Presented by ERIC P. WULFF

A candidate for the degree of

DOCTOR OF PHILOSOPHY OF LEARNING, TEACHING, AND CURRICULUM

EMPHASIS IN SCIENCE EDCUATION

And hereby certify that, in their opinion, it is worthy of acceptance

_____

TROY D. SADLER, Ph.D.

_____

PATRICIA J. FRIEDRICHSEN, Ph.D.

_____

BENJAMIN HERMAN, Ph.D.

_____

REX COCROFT, Ph.D.

Acknowledgments

Competition of the doctoral degree has been an experience of great wonder and exploration. I reflecting of the work that I have done in completing this dissertation, I would like to take this opportunity to thank the individuals that have helped me along my journey.

First, I would like to thank both Dr. Troy Sadler and Dr. Patricia Friedrichsen for serving as co-advisers, and providing me with extensive feedback and suggestions that helped improve the quality of my dissertation. I understand that this required extended periods of time, and I am grateful for their assistance.

I would also like to thank Dr. Ben Herman, and Dr. Rex Cocroft for serving on the dissertation committee. Ben provided me with assistance and guidance with regard to thinking through my statistical analysis, and Rex provided a grounded perspective of my work that challenged me to focus on the larger impact of my dissertation work.

Finally, I would like to thank my friends and family that supported and encouraged me along the way. Without the support of everyone this work would not have been possible. Thanks for everything.

TABLE OF CONTENTS

Table of Figures

EXPLORING THE RELATIONSHIP BETWEEN STUDENTS' LEVEL OF CONTENT KNOWLEDGE AND THEIR ABILITY TO ENGAGE IN SCIENTIFIC ARGUMENTATION USING STRUCTURAL EQUATION MODELING

Eric Wulff

Dr. Patricia Friedrichsen & Dr. Troy Sadler , Co-Dissertation Supervisors

ABSTRACT

The release of the Next Generation Science Standards (NGSS) in 2013 introduced science standards that are rich in core ideas as well as science and engineering practices. The NGSS views science content and science practice as closely interconnected to each other. The aim of this study is to explore the relationship between students' level of science content and their ability to engage in the practice of scientific argumentation. Specifically, this study teases apart content knowledge into both domain-general and discipline specific knowledge. To this end, this study explores the following research questions. (1) What is the relationship between students' content knowledge and their ability to engage in scientific argumentation? (2) How do the different dimensions of argumentation vary in difficulty? To explore these research questions, factor analysis, Item Response Theory, and Structural Equation Modeling are used. The results indicate the there is a stronger relationship between discipline specific knowledge and argumentation. This study contributes to the understanding of the connection between content knowledge and argumentation has the potential to inform and improve argumentation instruction, which ultimately can provide students with more authentic science experiences in the classroom

Chapter 1: Introduction

The release of the Next Generation Science Standards (NGSS) in 2013 introduced science standards that are rich in core ideas as well as science and engineering practices. Specifically, the NGSS focuses on merging three different dimensions of science—disciplinary core ideas, science and engineering practices, and crosscutting concepts as a way to develop students' ability to explain science phenomena and solve problems (NGSS Lead States, 2013). The NGSS replaced the previous National Science Education Standards (NSES). The NSES were released in 1996 and they have been used by states to support the creation of their own science standards as well as various assessments of student proficiency in science (NRC, 1996). Despite the wide adoption of NSES, there were concerns in the science education community that the NSES was not reflective of current research around the ways that students learn science (NRC, 2012).

The NGSS has several fundamental differences from the NSES. The NSES focused largely on inquiry and core ideas that were represented as separate standards (Reiser, 2013). In contrast, the NGSS aimed to support and encourage students' ability to explain science phenomena and solve problems through an integration of science and engineering practices, disciplinary core ideas, and crosscutting concepts. The notion of supporting student learning through the integration or weaving of these three dimensions of science is quite new in standards for teaching science. Another important difference is the move away from inquiry, which in the NSES standards was a very broad and not well-defined term (Anderson, 2002).

In addressing this issue, the NGSS clarifies inquiry by identifying specific science and engineering practices (S&EP). In doing this, the NGSS have sought to use the

terminology of practices to provide an elaboration and clarification on what it means to engage in science inquiry and how that can ultimately support student learning in efforts to build scientific knowledge (Reiser, 2013). A key component of the Framework is the inclusion and use of the term practice, which in the Framework refers to "not only skill but also knowledge that is specific to each practice" (NRC, 2012, p. 30). The unique approach of the NGSS promotes development of the students' ability to explain scientific phenomena and design solutions to problems through their engagement with science practices. This support students simultaneously developing knowledge using disciplinary core ideas and crosscutting concepts. Disciplinary core ideas consist of specific science ideas that are related to discrete science topics such as natural selection or chemical reactions. There are a total of eight science and engineering practices that include, but are not limited to constructing explanations, designing solutions, and engaging in argumentation from evidence. There are seven crosscutting concepts including patterns, and energy and matter. The unique process of integrating these three dimensions together into the process of learning science should help students build rich networks and mental models of scientific ideas and scientific phenomena. The integrated nature of the standards are important as Krajcik, Codere, Dahsah, Bayer, & Mun, (2014) argue that as more connections are developed, it greatly increases the ability of students to solve problems, make decisions, and make sense of new information. The integration of these three dimensions is unique and important in order to authentically represent the process of how students learn about the world. Ultimately, the weaving of the three dimensions of the NGSS aim to support the acquisition of scientific literacy by students and their development of critical thinking skills that allow them to explain phenomena and design

solutions to problems. In the NGSS this type of integration is referred to as 3-D learning. To support 3-D learning, the science and engineering practices, the disciplinary core ideas and the crosscutting concepts should be "interwoven in every aspect of science education, most critically, curriculum, instruction, and assessment (Pellegrino, Wilson, Koenig, & Beatty, 2014). In summary, the NGSS perspective on science education places great emphasis on main ideas of science (DCI), and integrates them with science and engineering practices. Thus, the NGSS moves the field of science education to adopting a perspective on content and practice that sees these two entities as intimately connected. Operating from within this integrated perspective, this study will explore the relationship between student's ability to engage in scientific argumentation and their knowledge of water systems.

Framing Argumentation as an Integrative Practice

The practice of argumentation is central to science. An abundance of science education research supports the notion that argumentation should be a key feature of learning and teaching science (Driver, Newton, & Osborne, 2000; Erduran & Jimenez-Aleixandre, 2008; Khine, 2012). The authors of the new Framework (NRC, 2012) defend the importance of argumentation by pointing out its centrality to the nature of scientific enterprise. Argumentation is, at its core, how scientific theories gain acceptance. Scientists are constantly engaging in the defense of their ideas, a process that can happen in a variety of settings from informal lab meetings to the formal process of peer review (Latour & Woolgar, 1986). As stated by Crombie (1996), "The history of western science is the history of a vision and an argument" (p. 12). Furthermore, argumentation is seen as a useful practice for citizens engaged in everyday negotiation of scientific claims. As

noted in the Frameworks, "Becoming a critical consumer of science is fostered by opportunities to use critique and evaluation to judge the merits of any scientifically based argument" (NRC, 2012, p. 71). In short, scientific argumentation is seen in the science education community as an essential practice for both scientists and citizens alike.

This study will focus on the practice of engaging in argumentation from evidence as an integrative science practice in the sense that argumentation requires competency in other practices. Ford (2008) states that engaging in scientific argumentation requires higher order cognitive skills, especially when engaging in critique. The use of these higher order cognitive skills in argumentation also draws on other science and engineering practices. For instance, successfully engaging in argumentation requires students to align and select evidence to include in their argument that provides supports for the merits of their claim. The process of identifying supporting evidence ultimately requires students to have made sense of a data source, and analyzed data that can be incorporated in their argument as evidence, which parallels the NGSS science and engineering practice of analyzing and interpreting data. Furthermore, when constructing the warrant for their argument, students must integrate a scientifically accurate statement as a means to provide a linkage between their evidence and claim. This process involves students constructing an explanation for a scientific phenomenon, and using this generated explanation in their argument to justify the manner in which their evidence supports their claim. Constructing an explanation of a science phenomenon is another key NGSS science and engineering practice. Argumentation has been described as a social practice (Sampson, Enderle, Gleim, Grooms, Hester, Southerland, & Wilson, 2012), as many researchers have made the case the arguments are ultimately shared with others as

they are intended to convince others of the merits of the argument's claim. While arguments can be shared in a number of ways, regardless of the distribution, they represent an instance in which ideas are communicated. This social dimension of argumentation requires that students engaging in argumentation as well as critique, allow students to both communicate their ideas, and also evaluate the ideas of others. Such competencies are outlined in the NGSS science and engineering practice of obtaining, evaluating, and communicating information. In order for students to demonstrate proficiency in augmentation, they must successfully engage in other NGSS practices (i.e. analysis and interpretation of data, explanation of science phenomenon, evaluation and communication of ideas). I argue that the practice of argumentation is an integrative practice for the broad spectrum of NGSS practices.

Operationalizing Argumentation

The practice of argumentation involves making claims using evidence to support those claims, and constructing warrants or reasoning statements to provide a convincing and meaningful connection between the claim and evidence of the argument. Critique and rebuttals are integral as well. Much of this work draws on Toulmin's (1958) model of argumentation. Toulmin's model aligns with how argument is commonly used in science practice. Scientists generally begin with a claim (i.e. hypothesis) and test its legitimacy against empirically collected data. The Toulmin (1958) model begins with a claim as a "conclusion whose merits we are seeking to establish" (p.90). In order to establish the merits of a claim, one must use evidence for support. As a result, the connection between the evidence and the claim is provided by a warrant that forms the substance of the justification of the claim. Within his model for argumentation, Toulmin also describes

backing, which are the implicit assumptions that warrants are dependent upon. Furthermore, claims may also be bounded by the use of qualifiers, which denote the limits of the claim's validity. For the purposes of this work claims, evidence, and warrants (reasoning) are the essential components of complete arguments. As argued by Ford (2008), argumentation is used to justify the validity of explanatory hypothesis, experimental designs, and interpretations of data sets. To that end, critique is essential to identifying flaws in arguments. Thus, critique is a necessary element of argumentation. The ability to critique requires slightly different competencies from argumentation in general. For instance, rather than forming a claim, critique requires the ability to first identify what the claim is in a given argument, what factors comprise the data that are used in an argument, and what type of reasoning is used to provide a sufficient linkage between the argument's claim and evidence. Engaging in critique requires students to make conclusions about the sufficiency of the argument's evidence, coherence of the entire argument, and also consider the standards for how the evidence was collected.

Learning Progressions of Argumentation

Given the importance of the practice of augmentation to the practice of science, there has been work aimed to create a learning progression that reflects how students learn argumentation. Berland & McNeil (2010) proposed a learning progression of argumentation that focused on both students' work in argumentation as well as considered the instructional environment that would support student learning of the practice. This progression consists of multiple dimensions including the instructional context and the argument product. Using these dimensions, the learning progression characterized the ways in which students' arguments vary in complexity and

sophistication across grade levels and instructional contexts. A significant contribution of this work is that it provides evidence that argumentation ability develops over time, as demonstrated by the higher level of complexity seen in the 12th grade student's work. However, as noted by Osborne et al. (2016), the evidence used in constructing this progression was drawn from qualitative classroom observations, and the researchers were not able to systemically assess student competency with argumentation as defined by the levels of their learning progression. Such concerns highlight the importance of using properly developed assessment items to measure student competency at the various levels of the argumentation learning progression.

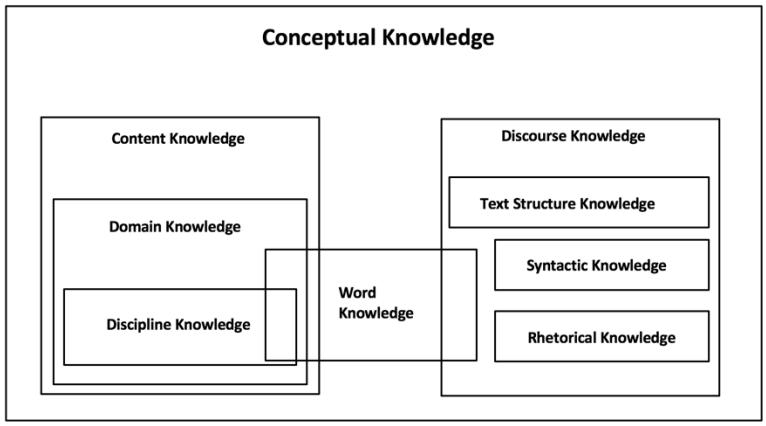
Osborne et al. (2016) also described a three-tiered learning of argumentation that accounts for the cognitive load placed on students as they engage in argumentation. One of the novel features of this progression is that it draws out the important distinction between argument construction and critique. A main goal of this study is to extend the work done to explore how these three tiers of the progression vary in difficulty as determined by relative difficulties of assessment items written to measure competency at these levels.

Framing of Content Knowledge

There are several perspectives across various disciplines in relation to subject matter knowledge that have been discussed within the psychology literature. Most of these perspectives are grounded within the field of educational psychology, and, within the field, there has been confusion with regards to terminology used to describe different aspects of knowledge. Alexander, Pate, Kuilikowich, Farrell, & Wright, (1991) posit that a critical problem with knowledge terminology within the educational psychology

literature is that authors tend to use different terms to refer to the same construct. For example, domain-specific knowledge (Alexander, Pate, Kulikowich, Farrell, & Wright, 1989; Walker, 1987), content-specific knowledge (Peterson, 1988), and subject matter knowledge (Voss, Blais, Means, Greene, & Ahwesh, 1986) all signify knowledge about a specific field of study (i.e. biology). Other researchers use terms, such as topic knowledge, background knowledge, and prior knowledge (Holmes, 1993; Alexander et al., 1991) to represent what other scholars might label as domain knowledge.

Figure 1.1: Conceptual Knowledge plane described in the Alexander et al. (1991) framework for content knowledge terminology.



In response to the confusion surrounding terms that describe aspects of knowledge, Alexander et al. (1991) describe the development of a framework for knowledge terminology. This framework consists of two major planes that are interconnected: conceptual knowledge and metacognitive knowledge. Within the framework, metacognitive knowledge refers to knowledge about knowledge. Despite the fact that metacognition is one of the most frequently studied constructs, there exists confusion in the ways in which authors define the construct Alexander et al. (1991). One of the more widely used definitions of metacognition was proposed by Flavell, who defines metacognition as "a part of one's acquired world knowledge that has to do with

cognitive matters" (1987, p. 21). Flavell (1987) subdivided the construct of metacognition into three categories of variables (person, task, strategy), which helped inform the design of the metacognition plane in Alexander et al's (1991) framework for knowledge terminology. The other plane of Alexander's framework of knowledge terminology is focused on conceptual knowledge. Conceptual knowledge is framed as a large construct that represents an individual's knowledge of ideas and concepts, and is sub-divided into content knowledge and discourse knowledge. Although content knowledge and discourse knowledge are depicted as separate discrete entities, Alexander et al. (1991) notes that these constructs are developed and used interactively. In the framework both content and discourse knowledge are presented in a hierarchical format, with each construct subdivided into smaller categories.

Alexander et al. (1991) define content knowledge as the knowledge of some aspect of one's physical, social, or mental world that can be formally or informally acquired. The framework subdivides content knowledge into domain knowledge, and then further sub-divides domain knowledge into discipline knowledge. A key part of the conceptualization of content knowledge is that it is hierarchal in the sense that as a concept becomes part of a more specialized field of study, it becomes part of the subcategories of conceptual knowledge. With that nuance of conceptual knowledge, Alexander et al. (1991) defines domain knowledge as a formal subset of conceptual knowledge; knowledge of a specialized field of study. Discipline knowledge, often referred to in the literature as topic knowledge or topic specific knowledge, is defined as a formal subset of domain knowledge, which focuses on a specialized field of study or a specific academic topic of a single lesson (Alexander et al., 1991; Bazerman, 1985).

It is within this conceptual knowledge plane of the framework that parallels exist to the NGSS dimensions. First, the core ideas (DCI) described in the NGSS are situated within both domain and discipline knowledge. From an NGSS perspective, the DCIs represent significant science ideas, concepts, and processes that students learn. In the conceptual knowledge plane of the Alexander et al. (1991) framework, DCIs are represented within domain knowledge. Domain knowledge is situated within the broader construct of content knowledge, and it refers to the understanding of science concepts and processes that students develop, which directly relates to the DCIs. In regards to water systems content, domain knowledge would include student knowledge of general water systems processes that are reflected within DCIs. A major focus of the NGSS is on integration of practices, DCIs, and crosscutting concepts, and while all of these are included explicitly in the framework, students will need to draw on both content knowledge and what Alexander et al. (1991) refers to as discourse knowledge. Within the framework, discourse knowledge broadly refers to students' understanding of language and is activated when students read to write. Given that several of the NGSS practices include processes like analyzing data, developing arguments, and developing models, these tasks at some level require students to be engaged in the processes of reading and writing. In that sense, students use discourse knowledge when they are engaging in science practices.

The framework also includes word knowledge, student's vocabulary or lexicon, as overlapping all levels of content knowledge and discourse knowledge. This addition represents the idea that the reading level of the individual can influence the manner in which content knowledge is used. For example, if a student has a high level of

specialized knowledge about a particular topic (i.e. knowledge about water systems), they may still have difficulty accessing and using this knowledge if it required making sense and understanding terms that they are unfamiliar with. This can also be the case if students are not able to complete an assessment task because the task includes terms not within a student's lexicon. This has implications for the ways in which individuals are able to build and use their content knowledge (Graves, Slater, & White, 1989).

The focus of this research study is on the construct of content knowledge, and the subdivisions within it. In a general sense, content knowledge in this study refers to the facts, ideas, and principles that are generated through scientific processes and investigations. This study will explore students' understanding of water systems, as well as more specialized knowledge of water systems. For the purposes of this study the term content knowledge refers to general knowledge about earth science as a very broad discipline. Domain knowledge would then refer to student's knowledge about water systems in general, and discipline or topic knowledge would refer to knowledge about specific topics within water systems, such as topography, surface water, water pollution, condensation and evaporation, groundwater, and infiltration. This study will draw on the Alexander et al. (1991) framework to explore the connection of different types of knowledge and practice, with the intention of better understanding the extent to which content knowledge is required to successfully engage in practice, specifically argumentation.

Research Questions

This study seeks to explore the connection between content knowledge and argumentation. In exploring this relationship, this study focuses on how students' level of

content knowledge affects students' ability to engage in scientific argumentation. In studying this relationship, this study focuses on different levels of content knowledge and how both discipline knowledge and domain knowledge can influence argumentation. In addition to this relationship, this study also explores the argumentation construct with regards to analyzing the ways in which different dimensions (i.e. alignment of evidence, structure, critique) of argumentation vary in difficulty. In order to address these research topics, this study includes three main research questions.

Question 1a

What is the relationship between students' content knowledge and their ability to engage in scientific argumentation?

Question 1b

Does discipline knowledge impact students' abilities to engage in argumentation more than domain-level content knowledge?

Rationale

The main goal of this set of research questions is to explore the relationship between content knowledge and argumentation. Researchers have examined the extent to which specific content knowledge impacts students' ability to construct and critique arguments (Koslowski, 1996; Kuhn, 1991; Lawson, 2003), and these studies have concluded that student do need some level of content knowledge to engage in argumentation. Despite being a focus of research, the relationship between content knowledge and argumentation is not completely understood. Previous work has focused on content knowledge in a more general sense, focusing on domain level content knowledge and studying how that impacts students' ability to engage in argumentation

(Hogan & Maglienti, 2001; Lawson, 2003; Zohar & Nemet, 2002). In exploring the connection between content knowledge and argumentation, this study analyzes the impact of both domain general and discipline-specific knowledge on students' argumentation ability. Better understanding this relationship can answer questions about students who perform poorly on an argumentation task or assessment. Researchers have investigated the question of whether poor argumentation performance is a result of a lack of general competency or a lack of content knowledge (Hogan & Maglienti, 2001; Koslowski, 1996; Lawson, 2003), but the answer remains unclear. On the surface, it might be easy to suggest that if a student is not able to perform well on an assessment of argumentation, then his or her argumentation skills must not be sufficient. However, there may be more variables that contribute to a student's argumentation ability that should be taken into account. One of these variables is the student's understanding of science concepts and principles, which potentially can heavily influence a student's ability to argue about a topic. The findings of this study will contribute to our understanding of the connection between content knowledge and argumentation, which will allow researchers to better understand the factors that can influence students' abilities to engage on argumentation.

Question 2

How do the different dimensions of argumentation vary in difficulty?

Rationale

The main purpose of this research question is to better understand how the different dimensions of argumentation can vary in difficulty. Research into how students learn argumentation have led to the development of learning progressions for

argumentation. Osborne et al. (2016) proposes a learning progression for argumentation that includes various dimensions of argumentation such as argumentation structure, alignment of claim and evidence, and critique. While Osborne et al. (2016) proposes a progression of how students learn the various dimensions of argumentation, there have not been specific assessment items that measure student ability at each of these levels. A goal of this study is to determine if the dimensions of argumentation described by Osborne et al. (2016) vary in difficulty. Better understanding the relative difficulties of the various dimensions of argumentation allows for better assessment items to be created that can discriminate students' argumentation competencies.

Question 3

How does content knowledge impact student performance on assessment items that vary in difficulty based on an argumentation learning progression?

Rationale

This research question explores the connection between content knowledge and various dimensions of the argumentation construct. This allows for a deeper understanding of how content knowledge may be more useful for engaging in certain dimensions of argumentation. For instance, content knowledge may be more important for certain argumentation tasks (i.e. alignment of claims and evidence), and less critical for other dimensions. This question builds on the first research question by examining the difficulties associated with the various dimensions of argumentation. This understanding has the potential to inform assessments of argumentation, as items can be written to argumentation dimensions that vary in difficulty. For assessment designers, it is crucial to recognize the relative difficulty of the various dimensions within the construct that the

assessment measures. This study can provide insight into the relative difficulties of the dimensions within the argumentation construct.

Summary of Chapters

This chapter provides an introduction to the problem being studied and a rationale for the research questions. Chapter 2 provides a review of the literature relevant to this study. Chapter 3 describes the research design and detailed methodology that will be used to conduct this research. Chapter 4 describes the results of the research study, and Chapter 5 includes a discussion of the study, focusing on the implications of the results and directions for future research.

Chapter 2: Review of Literature

For this study, I draw on a broad literature base within the field of science education to understand the construct and dimensions of argumentation. This review also focuses on the construction of learning progressions within science, and explores learning progressions for argumentation. A main focus of this is study is the examination of the relationship between argumentation and content knowledge, thus this review also summarizes research pertaining to the use of content knowledge in argumentation within science education. This review is comprised of the following sections: (1) argumentation in science education, (2) assessments of argumentation in science education, (3) research exploring the relationship between content knowledge and argumentation, (4) study of learning progressions in science education, and (5) learning progressions of argumentation.

Argumentation in Science Education

Scientific argumentation has been identified in the new Framework for K-12 Science Education (NRC, 2012) and Next Generation Science Standards (NGSS) (NGSS Lead States, 2013) as a central scientific practice to be included within curriculum and assessment. A significant direction of research within the science education community supports the notion that argumentation should be a key feature of the learning and teaching of science (Driver, Newton, & Osborne, 2000; Erduran & Jimenez-Aleixandre, 2008). Argumentation is an integral part of how scientific theories gain acceptance. Scientists are constantly engaged in the defense of their ideas, a process which happens informally during collaboration, lab meetings, and formally, as part of the process of peer review (Latour & Woolgar, 1986). In addition to researchers, argumentation is also seen

as a worthwhile practice for citizens who are engaged in the everyday negotiation of scientific claims. As evident in the Framework, the use of critique and evaluation of the merits of a scientifically grounded argument is required to become an informed citizen and effective consumer of science (NRC, 2012). Thus, the ability to engage in scientific argumentation is important for both scientists and citizens.

Argumentation is a widely used construct within science education, and researchers view argumentation in different ways. Argumentation that is scientific in nature is often described as a form of "logical discourse whose goal is to tease out the relationship between ideas and evidence" (Duschl, Schwiengruber, & Shrouse, 2007, p.33). Others have described scientific argumentation as a knowledge building practice in which individuals propose, support, critique, and refine ideas in an attempt to make sense of the world and the scientific phenomena within it (Driver, Newton & Osborne, 2000). Such perspectives on argumentation place emphasis on the construction and refinement of ideas with appropriate evidence. Osborne, Henderson, and McPherson (2016) identify the main goal of an argument as persuasion. They described scientific argumentation as a complex form of reasoning that requires domain-specific knowledge to construct and critique claims and their relation to any supporting evidence (Erduran, Simon, & Osborne, 2004; Osborne et al., 2016).

Framing of argumentation within science education

Given the various perspectives on scientific argumentation and the consensus of its significance to science education, it is important to consider that ways in which scientific argumentation has been framed. Within the field of science education there are multiple perspectives on argumentation. Sampson and Clark (2008) defined an argument

as the artifacts that a student or group of students produce when asked to justify claims or explanations, and the process of constructing those artifacts. Building from that definition, Sampson and Clark outline three important issues for researchers who study argumentation. These issues include the structure and complexity of an argument, the content of an argument, and the nature of an argument's justification. The themes provide a lens by which to analyze student arguments, which include emphasis on the core components of arguments and the ways in which they are constructed.

Toulmin (1958) proposed a framework, Toulmin's Argumentation Pattern, for augmentation that describes the multiple components of arguments: claims, data, warrants, backing, qualifiers, and rebuttals. Toulmin's framework for argumentation structure has had wide appeal, and has been used by researchers to examine the quality of arguments both within and outside of science education. The Toulmin model begins with a *claim* described as a "conclusion whose merits we are seeking to establish" (p. 90). To establish the merit of a claim, one uses *data or evidence* that can provide support. In turn, the relationship between the evidence and the claim is provided by a *warrant* that forms the substance of the justification for the claim. In general, warrants are reliant on assumptions that Toulmin denoted as *backing*, and the legitimacy of claims may also be restricted by the use of *qualifiers* that define the limits of validity.

Ford (2008) argues that an argument is used to justify the validity of explanatory hypotheses, experimental designs, and interpretations of a given data set. The process of critique allows for the identification of flaws within various components of arguments. For example, a critique might expose faulty reasoning of an argument that fails to provide justification of the evidence used in the argument. The focus of critique on evaluation

makes it an essential component of the argumentation construct. The ability to critique requires somewhat different competencies than would be required for creating an argument. For instance, rather than constructing a claim, it requires the ability to identify what the claim is in an argument, what constitutes the data used to support the argument, and what kind of reasoning is used to connect the data to the claim. A critique, therefore, considers whether the reasoning appropriately provides justification for the connection of the data to the claim that the argument is advancing, which requires knowledge of the features of an argument. The same skill set is also required for forming a rebuttal that would explain why the reasoning in a given argument is flawed. Commonly, this requires the cognitive skills of comparing and contrasting the relative merits of multiple arguments simultaneously, which also may involve constructing an argument for why particular pieces of evidence may have higher epistemic validity (Ford, 2008). Critique pertains to the considerations that have to be made about the accuracy and legitimacy of the evidence that is used in an argument.

In addition to perspectives on argument structure and critique, it is also important to consider how to evaluate the ways students incorporate ideas into arguments. Zohar and Nemet (2002) developed a framework for evaluating the quality of written arguments. In designing their framework, argument was defined as either assertions or conclusions and their justifications. Zohar and Nemet (2002) argue, "Argumentation is a type of informal reasoning because it involves reasoning about causes and consequences and about advantages and disadvantages, or pros and cons, of particular propositions or decision alternatives" (p.38). From that perspective, it follows that good arguments must include strong justification. The use of the phrase "justification" by Zohar and Nemet is

consistent with what many other scholars have termed evidence and reasoning, in the sense that these justifications provide the needed support for the claim that the argument is advancing. Within their framework, Zohar and Nemet describe the different ways in which students incorporate science ideas into an argument. These include: no consideration of scientific knowledge, inaccurate scientific knowledge, nonspecific scientific knowledge, and correct scientific knowledge. From the varying perspectives on argumentation, many scholars have argued that the construct includes various components, and crucial to argumentation is the ability to critically evaluate the merits of competing arguments in light of evidence and the justification that the arguments contain.

Assessments of Argumentation in Science Education

Given the central importance of argumentation to science, there have been different approaches to assessing students' ability to engage in augmentation. These approaches range from classroom observations to written assessments. Within the range of written responses, there also exists great variety in the nature and scope of these assessment tasks. While most of these tasks employ a constructed-response design, there are a few assessments that feature multiple-choice items, with the intention of creating assessments that are easy to use and score, while also demonstrating sufficient test reliability.

Methods used to elicit Argumentation

Despite the fact that argumentation has been a primary focus of research in science education (Lee, Wu, & Tsai, 2009) and is at the core of the Next Generation Science Standards (NGSS Lead States, 2013), there is limited research around how to assess argumentation (Osborne et al., 2016). Perhaps some of the these difficulties are

due to the challenging nature of designing these types of assessments, as numerous challenges have confronted researchers who are interested in measuring argumentation competency. One challenge for assessment designers is differentiating between the components of student arguments using a framework. As mentioned earlier, there are multiple frameworks that have been used to provide a perspective on scientific argumentation, and, in some cases, it can be difficult for researchers to decide, what counts as evidence, data, and backing in arguments that students create (Jiménez-Aleixandre, Rodriguez, & Duschl, 2000). This may be a contributing factor to previous work that reported difficulty in capturing the nuances of student reasoning using the framework that was proposed by Toulmin (Jiménez-Aleixandre, Rodriguez, & Duschl, 2000).

Assessments of written argumentation responses

Many researchers have chosen to use written responses to assess the manner by which students engage in argumentation. Argumentation assessments that include written responses can vary greatly in terms of both product and target audience. Examples of written responses assessing argumentation can range from short sentence length responses to full-scale formal written term papers, with target audiences spanning from elementary school to university level learners.

Several of these assessments present students with data or evidence, and through a written response, the student is required to construct his or her own argument with appropriate evidence and reasoning. In some cases, the assessment provides scaffolding (Osborne, et al., 2016) in order to make the task more approachable for a younger audience. Sampson & Clark (2008) developed an assessment that asked students to

determine which explanation, among six plausible alternatives or one of their own design, was the most valid or acceptable way to explaining a set of observations made from using available data. Students were then asked to formulate a written argument that articulated and justified that explanation with appropriate evidence and reasoning. This written task requires students to make connections between a claim and available evidence. A successful response included appropriate justification that provided a rationale for the use of a particular piece of evidence that supported a claim.

While both of the assessments described by Sampson and Clark (2008) and Kelly and Takao (2002) required students to formulate written arguments, they did not include explicitly elements of critique. More recently, there have been written assessments that include elements of critique. A task developed by Osborne et al. (2016) presented students with contrasting arguments addressing a particular topic, and, through a written response, students are required to critique the provided arguments and/or construct their own arguments that they believe to be superior. This is similar to work done by Sampson et al. (2013) which yielded an assessment that presented students with an argument by an quasi-expert that used an explanation involving data that was flawed in a noticeable manner. Students were then asked to refute the expert's claim using information and data provided in the question, and also present and support a counterclaim using evidence and proper justification.

Analysis of classroom discourse

There have been a variety of assessments of students' argumentation competency that are based on classroom discourse even though analysis of classroom discourse proves to be a labor-intensive process. Analysis of classroom discourse does provide

researchers with an opportunity to observe students' argumentation competency as a social practice, with students engaging in dialogue and discussion with their peers. Chin & Osborne (2010) designed a task in which students work in small groups to construct an argument using a template sheet that was provided. The template sheet contained writing stems, which required students to state their claim, data or evidence, reason(s) for their answer, counter-argument, and rebuttal. To help students visualize their ideas diagrammatically, each group was also required to represent their argument in graphic form, which was based on Toulmin's (1958) structure of an argument. The authors analyzed both the arguments that students had written, as well as the their oral questions and discourse. The quality of oral arguments was assessed using an analytical framework that accounted for the level of justification that was used to support the claim and how well a rebuttal addressed the weaknesses in an opposing viewpoint. Such an approach allows students to both engage in argumentation using both oral and written formats.

The approach described by Chinn and Osborne (2010) is similar to the design used by McNeil and Pimentel (2010), which included a written response and a task that involved a discussion where students shared and defended their arguments. This design introduces the potential for critique, but the analysis used by McNeil and Pimentel focused primarily on claim, evidence, and reasoning using Toulmin's (1958) structure. While assessments of classroom discourse of argumentation allow students to engage in the social aspects of argumentation, it requires a more labor-intensive analysis than some of the other assessment methods.

Multiple-choice assessments of argumentation

In contrast to assessment instruments that include written responses and classroom discourse, there are fewer instruments that include multiple-choice questions to assess student argumentation competency. While these types of assessments are much easier to score, they often are difficult to design. Strimaitis et al. (2014) developed a two-tier multiple-choice instrument that comprised two articles selected from the popular media that contained claims of general interest to a broad audience. The two-tier items were designed so the first tier was a multiple-choice question with two or three choices that assessed the student's ability to evaluate aspects of the claim. The second tier was a multiple-choice question with four possible reasons that assessed the logic used to determine the response to the first tier. This assessment is primarily focused on measuring the competency of students to evaluate the merits of scientific claims. While this instrument is not explicitly framed in terms of argumentation, the authors note that it does provide a credible measure for assessing whether a college student's knowledge of science practices is sufficient to critically read a scientific media article outside of the classroom (Strimaltis et al., 2014).

Osborne et al. (2016) also described attempts to develop multiple-choice items to assess argumentation. Much of this work stems from the fact that multiple-choice items are attractive assessment options, since they can be scored reliably and cheaply, which is not true of the majority of assessments of scientific argumentation that analyze written responses or discourse. While the final assessment produced by Osborne et al. (2016) did not contain multiple-choice items, the authors did indicate that earlier iterations of the assessments included both multiple choice and constructed response items. The authors

reported that the multiple-choice items were frequently mis-fitting and/or artificially easy, so they were removed from the final version of the assessment. They also reported that the findings from multiple rounds of item analysis indicated that multiple-choice items testing argumentation do not elicit the same competencies as constructed response items. This is an area that needs to be explored further, but the lack of multiple-choice assessments makes this a difficult area to research.

Research exploring the relationship between content knowledge and argumentation

Argumentation has been the focus of a significant amount of research within science education. Specific research has examined the role that content knowledge plays in students' ability to construct and critique arguments (Kuhn, 1991; Lawson, 2003; Sadler & Donnelly, 2007) and has concluded that students need some level of content knowledge in order to engage in scientific argumentation. This may be over-simplistic as research on the development of student ideas suggests that this development does not follow a linear pattern (Steedle & Shavelson, 2009). However, further understanding the connection between argumentation and content knowledge allows researchers to not only understand how students generate arguments, but it also allows for an understanding of the ways that students use their pre-existing knowledge in new ways. The goal of this section of the chapter is to investigate the extent to which students require science content knowledge in order to engage in argumentation involving science ideas.

Zohar and Nemet (2002) investigated the effects of a genetics unit that incorporated argumentation instruction on students' knowledge of genetics, argumentation competency, and the integration of genetics knowledge into an argument. The authors used four categories to rate arguments with respect to the level of content

integration: no biological knowledge considered, incorrect consideration of biological knowledge, consideration of non-specific biological knowledge, and correct consideration of specific biological content knowledge. Within the study, an experimental group participated in a unit that incorporated instruction in argumentation, while the control group was exposed to a curriculum that covered genetics in a more traditional format, which did not include argumentation. The curriculum used in the treatment group was designed around recommendations in the literature (Kuhn, 1991; Voss & Means, 1991), which included explicit instruction about the formal structure of an argument, multiple opportunities for students to engage in argumentation, and taking part in elaborate discussions that required the use of argumentation. The researchers used a pre/post design in which students completed a written assessment task. When comparing the treatment groups, the authors reported that the experimental group performed better at integrating content knowledge into their arguments. The number of students that correctly incorporated specific science content was higher in the experimental group (53%) when compared to the control group (9%). These results indicate that argumentation-based instruction contributed to students' awareness of the importance of science content knowledge in constructing arguments.

There has also been work done by Sadler and others investigating how students' prior understanding of science impacts the quality of socioscientific arguments. Sadler and Zeidler (2005) examined the effect of college students' varying levels of content knowledge on the quality of their arguments. The researchers administered a quantitative test of genetics knowledge to two hundred and sixty college students with varying academic backgrounds, and selected a subsample based on the test results to conduct

interviews that were designed to assess argumentation in the context of genetic engineering. Sadler and Zeidler used a rubric based on Toulmin (1958) and Kuhn (1991) to assess the quality of the arguments that were reflected within the student interviews. Students who exhibited higher content knowledge, as measured by the genetics test, exhibited fewer reasoning flaws. Also, the students with high content knowledge often revealed their knowledge throughout the interview, while students with low content knowledge test scores frequently admitted their lack of knowledge and did not show evidence of applying their science knowledge in their arguments. The findings of this study suggest that there is a relationship between content knowledge and socioscientific argumentation. The higher the level of student content knowledge meant that they were able to utilize that knowledge within their arguments and commit fewer reasoning errors.

Sadler and Fowler (2006) further explored the ways in which students use science content with socioscientific arguments. Their study specifically looks at how students used science content to justify claims about genetic engineering. Within the study, interviews were conducted with 45 participants in three groups: high school students with variable genetics knowledge, college non-science majors with little genetics knowledge, and college science majors with advanced genetics knowledge. During the interviews, students advanced claims pertaining to various scenarios relating to cloning and gene therapy. Assessments of arguments included the number and quality of justifications using a five-point rubric. The study reported that college majors in science performed better than others in terms of the quality and frequency of justifications. Argumentation did not differ between high school students and nonscience majors. Follow-up qualitative analyses of interview responses indicated that all three groups had a tendency to focus on

similar, social-moral themes as they struggled with genetic engineering issues that had a layer of social complexity, but the science majors frequently referenced specific science content knowledge within their claim justification.

Sadler and Fowler's (2006) results support a threshold model of content knowledge transfer. This model proposes thresholds around which argumentation quality would expect to increase. The authors conclude that students require some level of content knowledge in order to engage in the task, and those students with advanced science content knowledge constructed better arguments than students with basic knowledge. In the model the authors describe a content knowledge scale that ranges from zero to three. Zero represents little or no content knowledge, level one refers to basic knowledge that has been described by Perkins & Salomon (1989) as the "rules of the game" which suggests that students would struggle to engage in meaningful argumentation without understanding of the content, level two is advanced knowledge, and level three refers to the knowledge expected of a professional or expert. The described relationship between argumentation quality and content knowledge based on this threshold model suggests that this relationship is not linear as there is no relationship within the thresholds.

The work mentioned thus far used various scenarios to elicit argumentation to examine how content knowledge interacts with argumentation. Overall, the findings indicate that the relationship between the two is non-linear. While content knowledge is important for argumentation, there are other types of knowledge that may be relevant. Specifically, epistemic knowledge refers to knowing how we know what we know, and the manner by which scientific knowledge is generated (Hogan & Maglienti, 2001;

Hewson, 1985; Metcalfe & Shimamura, 1994). Thus, analyzing the content knowledge that students incorporate into an argument may not convey the entire story, as there are also epistemic considerations.

Hogan and Maglienti (2001) investigated the role of epistemic knowledge on constructing argument critiques by interviewing scientists, adult non-scientists, and middle school students to determine how scientists and non-scientists construct arguments. Participants were asked to rate the validity of a set of ten conclusions based on empirical observations. The context of the task was an ecological problem involving purple loosestrife, which is an invasive plant introduced in the eastern Unites States from Europe. Participants were asked to rate conclusions that ranged from notions that the invasive species improves the ecosystem by adding beauty to detrimental effects on reasoning tied to empirical observations. For data analysis, the researchers used a holistic rubric that included levels 0-4 for the rating the participants gave the conclusions. The researchers took into account whether the participants perceived strengths and weaknesses of the provided conclusions were grounded in empirical findings. From the comparison of scores, the authors reported that students and adult non-scientists were similar, and they were both different from scientists. The authors also reported that students had a tendency to form their own conclusions after reading the evidence and then rated particular conclusions high if they were consistent with their own conclusions. Scientists, on the other hand, were more careful to note that it was not possible to assert any causal connections based solely on these observations. Based on this study, one may conclude that epistemic knowledge is a key component of argument quality since the authors reported that students looked to see if the potential conclusions were consistent

with their conclusions and the scientists looked to see if the conclusions were consistent with the empirical evidence. It is also important to point out that scientists had a higher level of content knowledge than the others in the study, but it seems that scientists were able to draw on epistemic knowledge when evaluating the conclusions more efficiently than the others.

Impact of Argumentation on Students' Conceptual Understanding of Science

Numerous researchers have explored the relationship between content knowledge and argumentation, trying to determine to what extent content knowledge is required in order for students to engage in argumentation. Additional work has focused on examining the potential for argumentation to support gains in science content knowledge. Several empirical studies have reported the positive impact of argumentation on students' understanding of scientific ideas and processes (Ayadeniz, Pabuccu, Cetin, & Kaya, 2012; Cross, Taasoobshirazi, Hendricks, & Hickey, 2008; Zohar & Nemet, 2002). There have also been studies that reported positive impacts of argumentation on epistemic engagement (Kelly & Takao, 2002) and the level of learning for students (von Aufschnaiter, Erduran, Osborne and Simon, 2008).

Cross et al. (2008) conducted a study with middle and high school students that involved using argumentation to teach a two-week unit that targeted science concepts. The instruction included the NASA Classroom of the Future's Bio BLAST instructional unit designed to engage middle and high school students in the collaborative learning of core biological concepts. The authors employed a case study design to study a group of students who engaged in argumentation within the context of the unit. Students were given daily quizzes that consisted of three to four 2-part items. The first part included a

short answer response and the second part required students to provide a justification for their answers. Quizzes were not graded, but rather each student was paired with a peer and engaged in structured argumentation to discuss their answers to the quiz questions. The results of the study indicated that engaging in argumentation resulted in greater understanding of pre-existing concepts and the exposure of students to new ideas by virtue of their peers helped them to extend their pre-existing knowledge. The authors also reported that this process has the potential to eliminate misconceptions, but more work needs to be done to substantiate that claim.

While the majority of studies report argumentation has a positive impact of students' conceptual understanding of science, other studies report conflicting results. Walker, Sampson, Grooms, Anderson and Zimmerman (2010) conducted an argumentation study with a group of college students that used Argument Driven Inquiry (ADI) in the lab sections of a chemistry course. The researchers compared the learning gains on the Chemistry Concept Inventory (CCI) between a group of students who experienced argumentation instruction (ADI) and a comparison group who received traditional instruction. The results indicate that the students who received ADI did not perform significantly better than students in the comparison group on a test that measured conceptual understanding of key chemistry topics, but the ADI group did outperform the comparison group on tasks that involve reasoning and using evidence. One important limitation of the study was the difference in the number of experiments that each group completed. The ADI group only completed six experiments, while the comparison group completed 11 investigations. Exposure to a fewer number of experiments may have

limited the content understanding that was measured by the CCI and may account for lower scores from the ADI group.

When looking broadly at the research on argumentation and content knowledge, argumentation may have a significant impact of students' learning of key science concepts in some contexts. Thus, it is important for science educators to explore the impact of argumentation on those science concepts that can be difficult for students to learn. Ayadeniz et al. (2012) designed a study to explore the impact of argumentation-based pedagogy on college students' understanding of the properties and behaviors of gases. The authors implemented argumentation-based pedagogy using Toulmin's (1958) model of argumentation and allowed students to create their own arguments throughout the instruction, while providing coaching on how to back up findings with credible evidence. A comparison group of students received a more traditional lecture-based curriculum. Reported findings indicated that the argumentation group performed better than the comparison group on an instrument that measured students' conceptual understanding of the properties and behaviors of gases. The authors argued that. in their study, argumentation created a context for learners to elaborate on pre-existing ideas; the process of writing arguments allowed students to organize their pre-existing ideas and communicate them in a convincing and coherent way. The authors also argue that engaging students in verbal argumentation, after they developed written arguments, may have helped students address the gaps in their own knowledge by listening to the ideas of their peers and asking questions to clarify their own understandings.

Learning Progressions in Science Education

Science learning progressions are generally defined as descriptions of the increasingly sophisticated ways that learners can think about a science topic over a period of time (Duschl, Schweingruber, & Shouse, 2007). Within the science education, learning progressions are viewed as valuable educational resources (Duncan & Gotwals, 2015; Smith & Weiser, 2015). Researchers have praised learning progressions as a potential guide for curriculum development that moves learners towards more sophisticated thinking in both disciplinary practice and content knowledge (Berland & McNeill, 2010; Songer, Kelcey, & Gotwals, 2009). Researchers have also argued that learning progressions are a valuable resource for the development of meaningful assessments (Alonzo & Steedle, 2008; Berland & McNeill, 2010). In short, the science education research community is exploring learning progressions because they have the potential to allow educators to coordinate curriculum demands, instruction, and assessment in a more effective way (Duschl et al., 2007).

Despite the agreed upon benefits of learning progression, many researchers have criticized the way researchers have developed learning progressions (Duncan & Gotwals, 2015; Duschl et al., 2011; Ford, 2015; Hammer & Sikorski, 2015; Shavelson, 2009; Smith & Wiser, 2015). For instance, Shavelson (2009) cautioned that learning progression research is susceptible to "data fitting" as research may ignore individuality in learner thinking. Duschl, Maeng, and Sezen (2011) identified a "flurry of competing perspectives" on learning progressions. Currently, different researchers are using the construct of learning progressions in different ways (Berland & McNeill, 2010; Duschl et al., 2011). Duschl et al. (2011) distinguished between evolutionary and validation-

learning progressions that differ in the ways that researchers viewed conceptual change. Duschl et al. (2011) views evolutionary forms of learning progression research as attending to the development of foundational knowledge, while viewing validation forms of research (i.e. developing assessment models, testing discourse strategies, instructional interventions) as components of science learning and thus are perhaps better labeled as teaching sequences than learning progressions. Additionally, Shavelson (2009) distinguished between curriculum and instruction learning progressions and cognition and instruction learning progressions. Shavelson explained that researchers begin with a logical analysis of science topics when developing curriculum and instruction learning progressions, while they begin with a psychological analysis of cognition when developing a cognition and instruction learning progression. In contrast to these perspectives, Berland and McNeil (2010) explained that some researchers use learning progressions as "developmental progressions" while others use learning progression as descriptions of levels of complexity of scientific knowledge and practice. A key distinction between these perspectives is that the latter places emphasis on the role of instruction (Berland & McNeil, 2010).

As noted above, there are a variety of perspectives that researchers take in studying learning progressions. Inevitably, the differences in the framing of learning progressions have implications on the various learning progressions that are developed. In general, the process of developing a learning progression begins with researchers identifying core science ideas that the progression will address (Duschl et al., 2011). Next, researchers identify a possible sequence of ideas that could logically lead to an understanding of the specific topic. This process often involves drawing upon current

standards for teaching the topic, and previous research about how students learn the topic. After creating an initial progression, researchers then work on refining and validating the progression using data on student thinking. The final step is an on-going process of iterative cycles often involving messy data and difficult decision-making (Shea & Duncan, 2013), which leads to an iterative process of refining the various stages of the progression based upon student data.

Alonzo and Steele (2008) described a related, but more specific, process for developing a learning progression. They explained that the learning progression development process starts with expectations about what learners should know about a concept, which often is based upon standard documentation or prior research about how students learn. These expectations about learner ability represent the upper end of a progression. Lower levels of the progression can be based on research about student ideas about the construct. Often, these ideas can be misconceptions that students may have about the construct. Alonzo and Steele grouped learner ideas based on similarities, and then these groupings were ordered in a logical fashion to create a hypothetical learning progression. This hypothetical progression represents a current idea about how student understanding advances and this hypothetical progression should be revised as new data is analyzed. After creating the hypothetical progression, the researchers can develop items to assess students' level of achievement. Data from assessments can then be used to inform both the assessment instrument and the learning progression.

Learning progressions of Content Knowledge and Science Practices

Research on learning progressions is an active area of research in science education. With this widespread popularity, there are several examples of learning

progressions that relate to science content knowledge. Among these various progressions, I describe a learning progression that pertains to water systems, since this knowledge is relevant to this study. Gunckel, Covitt, Salinas, and Anderson (2012) described an approach to developing learning progressions that was quite similar to the approach taken by Mohan, Chen, and Anderson (2009) and Jin and Anderson (2012). Gunckel et al. (2012) explained that their learning progression for water was developed and refined through iterative cycles of assessment and analysis over a 6-year period. The authors began with hypothesized lower and upper anchors, which were used to develop assessment items that elicit student thinking. Based on the assessment data, the authors developed intermediate levels of the progression in an iterative way throughout the design process. During the process, Gunckel et al. (2012) created "exemplar workbooks" that represented clusters of student ideas that "could be used to distinguish between qualitatively different patterns in student accounts" (p.852). Each exemplar workbook came to represent a specific level on the learning progression. These workbooks not only helped inform the design of the learning progression, but were used as tools to assign learners to levels of achievement. Regardless of the researcher's process for developing a learning progression, all researchers must begin with the same essential decision. They must identify a science topic that is worthwhile in the context of learning progression research. Generally, learning progression research has defined worthwhile topics as those that are considered central to scientific disciplines (Duschl et al., 2011; Gunckel et al., 2012; Jin & Anderson, 2012).

Gunckel et al. (2012) developed a learning progression for water in socio-ecological systems. Their progression includes four levels of achievement, using five

different elements. These elements include structures and systems, scale, scientific principles, representations, and human agency. The lowest level, *force-dynamic accounts*, include a human centric focus where water is identified in visible familiar contexts. Often these include lakes, rivers, oceans, and water is generally referenced as part of everyday life, as something that people can use. The next level, *force dynamic accounts with mechanisms*, shows an expanded awareness with the physical world. Significantly, this level includes mechanisms that can be used to move or change water. Level 2 accounts explain water movement or changes in water quality in terms of perceived natural tendencies of water or in terms of other living and non-living components acting on the water (i.e. clouds), but fails to describe hidden systems (i.e. groundwater) in detail or identify connections to invisible parts of the system such as atmospheric water. The third level, *incomplete school science accounts*, includes accounts that are characterized by the retelling of stories or parts of stories about the water cycle that are typically learned in school. Often these stories put events in order, including multiple pathways for water, and provide scientific names of processes that move water at the macroscopic scale. Despite this, these accounts are incomplete. These stories are missing steps or include errors in tracing water and substances in water systems. Furthermore, level 3 accounts do not include driving forces or constraining factors when tracing water or substances in water. The highest level of the learning progression is *qualitative model-based accounts*. At this level, accounts include the constraints that limit the capacity of environmental systems to provide freshwater. Importantly, these accounts identify driving forces and also consider the effect of constraining factors on the pathways for water and substances in water. The intermediate levels in the progression represent benchmarks in student achievement as

students progress from force-dynamic to scientific reasoning, Gunckel et al. (2012) derived the intermediate levels of student achievement from theories about how knowledge and practice are organized, and from empirical research (i.e. Briggs, Alonzo, Schwab, & Wilson, 2006).

Since the wide adoption of the NGSS, there has been a shift in science education to move beyond focusing on science knowledge exclusively, and include an additional focus on practice. However, despite this shift the development of learning progressions has not included science content knowledge specifically. Some learning progressions have focused on the development of conceptual knowledge without consideration of the scientific practices involved in constructing that knowledge. An example of this is the work of Plummer & Krajcik (2010) that describes a learning progression for celestial motion that does not include considerations for scientific practices. This lack of practice is also found in Alonzo & Steedle's (2008) learning progression of force and motion also does not include consideration of practices into the progression.

Sikorski and Hammer (2010) make the argument that learning progressions that focus mainly on student attainment of correct ideas of science may impede the ability of students to develop an understanding of science practices. The authors support this claim by suggesting that rather than using available evidence to assess ideas, students may instead focus on comparing the alignment of their ideas against recognized science knowledge. Such learning is in conflict with current understandings among many science educators, which have moved towards a focus on students' engagement in practices even if they involve non-canonical accounts of the phenomenon (Sikorski & Hammer, 2010). In contrast to learning progressions that do not include practices, there have been several

learning progressions that have included scientific practice. Lehrer and Schauble (2012) integrated a variety of scientific practices in their progression of evolutionary theory. For instance, their learning progression level 4A of the ecosystems strand involves asking questions, and level 4B involves modeling. Each of these practices is fully contextualized in students' understandings of conceptual knowledge of ecosystems.

In comparison, the focus of some learning progressions is on developing an understanding of scientific practices without explicit connections to content knowledge. Schwartz et al. (2009) discussed a learning progression on scientific modeling that was removed from a specific content area. The authors discussed the importance of integrating modeling with the meta-knowledge learners have about modeling. However, the authors ignored considerations about the conceptual context in which students are developing their understandings about modeling. Schwartz et al. (2009) admit, "The influence that specific contexts have on learning scientific practices is, of course, critical" (p. 635). The authors were wondered if the students demonstrated lower modeling ability because they held less sophisticated view about the practice of modeling or because they lacked content knowledge.

Similar to the learning progression for modeling, Berland and McNeil (2010) developed a generalized leaning progression for the practice of argumentation. In developing the progression, the authors recognized instructional context as an important dimension of argumentation, but contextual knowledge was not the focus of instruction. As a result, this learning progression of argumentation does not explore the role that content knowledge had on the ability for students to engage in the practice of argumentation.

Songer et al. (2009) took a different approach in designing their learning progression. Differing from progressions that include both content and practice, Songer et al. constructed two separate but parallel learning progressions of biodiversity. One focused exclusively on content knowledge, while the other on inquiry reasoning. The authors highlighted the importance of integrating the content of biodiversity with inquiry reasoning knowledge, so they decided to emphasize this integration in both of their learning progressions. So they emphasized that all of their learning progressions reference both components of the learning progression. Additionally, the authors explained: "In an ideal curricular unit manifested from our progressions, students could be working with one level of the inquiry reasoning progression (e.g., intermediate) many times in combination with different focal points along the content progression (p. 613)". Thus, the authors suggested that learning about science concepts is fundamentally different, even if they are related, than learning about science practices. Such an assertion justifies the recent focus of work on learning progressions for science practices.

The Berkeley Evaluation and Assessment Research (BEAR) Assessment System has served as a tool to aid in the development of learning progressions. The BEAR Assessment System is an approach to assessment development that provides meaningful interpretations of student work relative to the cognitive and developmental curricular goals (Wilson & Sloan, 2000). The BEAR Assessment System is grounded in four key principles that guide assessment development. These include the following: (1) assessment should be based on a developmental perspective of student learning, (2) alignment between instruction and assessment, (3) teachers are the managers and users of assessment data, and (4) classroom assessments must uphold traditional standards for

validity and reliability. These four principles relate to the Assessment Triangle developed by the National Research Council Committee in their report, *Knowing What Students Know* (Pellegrino, Chudowsky & Glaser, 2001). The Assessment Triangle is a model of the essential connections present in an assessment system. In the Triangle, assessment activities (observation) must be aligned with knowledge and cognition processes (cognition) that can be affected by instructional means, and the scoring of interpretation of student work must reflect measures of the same knowledge and cognitive processes (Pellegrino et al., 2001).

The BEAR Assessment System has been used to develop learning progressions by providing a system for aligning assessment items to the construct that is the subject of a learning progression. An integral part of this process includes the development of construct maps. A construct map defines a latent trait that is used to represent a cognitive theory of learning that is consistent with a developmental perspective (Wilson & Draney, 2004; Wilson & Sloan, 2000). This approach allows assessments that can determine how students are progressing on a scale from novice to expert grounded in a specific construct, rather than using assessments to measure competency after learning activities have been completed. To that end, the construct map includes different levels of knowledge that can be expected for students learning a construct, and ranks them by increasing levels of the sophistication of their understandings (Wilson, 2009).

An important piece of the BEAR Assessment System is designing items and tasks that elicit student knowledge in ways that are described in the construct maps. Wilson (2009) argues that a guiding principle in the assessment design process is the seamless integration of assessment to an instructional unit or sequence. Furthermore, items should

be written with the intention of producing evidence of specific levels of understanding along a construct. Often this will involve creating items that map to different levels in the construct map to allow for the instrument as a whole to estimate the level of student ability on the construct map. In some cases this can involve including distractor items that map to various levels of sophistication reflected in the construct map (Alonzo & Steedle, 2008; Briggs & Alonzo, 2012) but this has not been widely used with the BEAR Assessment System. Such an inclusion allows for multiple-choice items to provide more evidence for the level of sophistication in students' understandings of the construct than if the items were simply dichotomously scored.

In addition to item construction, use of the BEAR Assessment System also involves an appropriate measurement model that defines how inferences about student knowledge can be made from the assessment scores. This interpretation commonly involves the use of a psychometric approach that is based on the Rasch-based item response model known as the multidimensional random coefficients multinomial logit model (Adams, Wilson & Wang, 1997). This model provides a way of estimating student ability levels as well as item difficulties using the same scale. The model also allows for this information to be displayed graphically. This graphical plot, referred to as a Wright Map, plots student location on the left side (from low ability on bottom to high ability on top), and item difficulty on the right side (from easy on bottom to difficult on top). In interpreting the Wright Map, if student ability and item difficulty are close to each other, then a student at that ability level should have a 50% chance of getting the item correct. Likewise, if the item is at a lower level of difficulty than the student's ability level, that student will have a greater than 50% chance of getting that item correct. By use of

student response patterns and the difficulties of the items, the level of sophistication of students' ideas can be estimated. This estimation places students on the construct map that was used to inform the assessment design.

Learning progression of scientific argumentation

With a main focus of research and development of learning progression of science practices, there have been several researchers who have developed learning progressions that detail how students learn scientific argumentation. Berland and McNeil (2010) describe a learning progression for argumentation that consists of three dimensions: instructional context, argument product, and argument process that demonstrate the way that students' arguments vary in complexity and sophistication across grade levels. The context of the argument must be rich enough to allow for multiple perspectives (DeVries, Lund & Michael, 2002) and must require the use of evidence to reconcile multiple perspectives. This dimension of the progression also includes a consideration for the size and appropriateness of the data set used to inform an argument. Another consideration within the context of an argument is the level of scaffolds available. Berland and McNeil (2010) defined scaffolds as "temporary supporting structures provided by tools or individuals to support student learning of complex problem solving" (p.771). Scaffolds may be provided in a variety of formats from curricular materials to visual representations on a classroom wall. Scaffolds can help make the implicit rules of argumentation explicit for students, and can also simplify a complex task to make it more accessible to students. The second dimension of the learning progression focuses on the argument product. A complete argument product should include a claim that is defended with both evidence and reasoning, and the progression also accounts for the quality of

each of these components. This involves a careful consideration of the sufficiency and appropriateness of the evidence, reasoning, and rebuttal. Appropriateness answers the question if the evidence, reasoning, and/or rebuttal are relevant to the problem and scientifically accurate to support the claim a student is making, while sufficiency refers to the quantity or complexity of the evidence, reasoning, and/or rebuttal being able to convince an audience of the claim (Berland & McNeil, 2010). The final dimension of the learning progression, argumentation process, includes consideration for the manner in which arguments are articulated, defended, and revised. At high levels of the learning progression, the authors expect students to revise their ideas in light of compelling competing viewpoints. This is seen as the highest level of process, while at lower levels, students are able to recognize other competing viewpoints, but they are not revising their arguments in light of them. With this learning progression, Berland and McNeil (2010) found a range of complexity from 5th grade to 12th grade students.

There has also been work conducted by Songer, Gotwals, and colleagues that has investigated the development of students' thinking about complex ideas in biological science (Gotwals & Songer, 2010). They have also described a learning progression for how students construct evidence-based explanations (Songer, Kelcey, & Gotwals, 2009). Despite referring to their progression variable in terms of an evidence-based explanation, it essentially involves constructing arguments from evidence, which focuses on students' ability to advance a claim, support it with evidence, and use reasoning to provide a linkage between the evidence and claim. In their progression for evidence-base explanations, Songer, Kelcey, and Gotwals (2009) include three main levels: complex, intermediate, and minimal. This progression is simpler than the multidimensional

learning progression described by Berland and McNeil (2010), but it does include similar components. At the highest level of evidence–based explanation, students are able to construct explanations using claim, evidence, and reasoning, without any prompts or guidance. The inclusion of a consideration of prompts or guidance is also reflected in the Berland and McNeil progression, where it is referenced as scaffolding. Despite framing the learning progression in terms of argumentation, it is describing the same practice albeit with different terminology.

Lee, Liu, Pallant, Roohr, Pryputniewicz, and Buck (2014) make the case for using "level of uncertainly" as an important component of argumentation. The authors also advance a learning progression of argumentation, which consists of making a claim at the lowest level, and including higher levels all the way up to providing justification for any uncertainly. A significant contribution of this work is that it includes elements of epistemic knowledge within a learning progression of argumentation, which has been lacking in previous progression literature. The notion of uncertainty includes considerations that students must make about the evidence that they are using in their arguments, as well as applying the notion of uncertainly to evaluating other arguments. The authors make the case that uncertainty is important in the critique process.

Finally, Osborne et al. (2016) developed a learning progression that draws on the Toulmin (1958) model of argumentation. The proposed learning progression includes three levels, each with levels requiring increasing numbers of connections to be made between claims and evidence. The lowest level of the progression, denoted as zero, does not require explicit connections between claim and evidence to be made, and this level is termed zero degrees of coordination. The inclusion of a warrant to provide a linkage of

the claim and evidence marks the transition to level one. Osborne et al. (2016) articulate that level one of the learning progression "requires understanding of not only what constitutes a claim or a piece of evidence, but also how to construct or critique a relationship between claim and evidence" (p.827). As noted by the authors, students at this level are dealing with the cognitive demands of drawing connections between claims and evidence. At the most advanced level of the learning progression, students are expected to compare two or more warrants, and this level is referred to as two degrees of coordination. At this level, arguments become complex as the claims draw on multiple warrants that each draw on evidence. At this level, students are engaged in critiquing other arguments, which involves analysis of the accuracy and appropriateness of both warrants and evidence. This dissertation is an extension of this learning progression of argumentation by exploring multiple-choice assessment items at different levels of the progression. The findings of this study have the potential to inform the development of multiple-choice assessment items.

Summary

The survey of literature presented in this chapter was intended to provide an overview of the work that has been done in science education around argumentation assessment, integration of content knowledge into argumentation, and the use of learning progressions in argumentation. When reviewing the literature, it became apparent that there are areas of research that require further work and investigation. With regards to the connection of content knowledge and argumentation, while there has been research focusing on how content at a more general domain-level can influence students' ability to engage in argumentation, previous work has not explored how more specific discipline

knowledge can impact students' ability to engage in argumentation. Second, while there are several different learning progressions for scientific argumentation that have been developed, many of these progressions do not include specific assessments items for the various proficiency levels. This study will help inform both of these points, as it will explore the relationship of more specialized discipline level knowledge and students' argumentation ability. Analyses within the study will also explore multiple-choice item difficulties of different items at different levels of an argumentation learning progression.

Chapter 3 Research Methodology

The purpose of this study is to examine the relationship between content knowledge and argumentation. In exploring this relationship, this study also seeks to explore the dimensions of argumentation by analyzing assessment items that relate to varying levels on the Osborne et al. (2016) learning progression of scientific argumentation. Additionally, this study seeks to explore the extent to which content knowledge influences student ability to engage in the different dimensions of argumentation. This chapter serves to communicate the research methodology used to explore these issues. In this process, the chapter will summarize the sample population that was used within the study, outline the data collection procedures, and identify the data analysis techniques that were used to answer the research questions.

Description of Sample

This study collected assessment data from 827 middle school students across 10 mid-western middle schools. The data for this study are a subset of a larger dataset collected to investigate the impact of an argumentation-centered water systems curriculum on students' ability to engage in argumentation and also learn critical water systems topics. Due to the participation within the larger investigation, students who were included in this study had completed a 10-day instructional period that focused specifically on water systems. Within the sample, 632 students had completed the Mission HydroSci (MHS) curriculum, which in addition to covering water systems content also placed a heavy emphasis on scientific argumentation. MHS is a game-based curriculum that has been designed to teach major topics within water systems. These topics include: surface water, water pollution, groundwater, and atmospheric water. In

addition to teaching students about water systems content, the MHS curriculum also focuses on building students' argumentation skills. Within each unit, students are given the opportunity to formulate an argument by selecting appropriate claims, reasoning and evidence. Within the game, students collect data about the environment and use their collected data to support their arguments. Throughout the game, players argued about a range of topics from where to set up base camp based on the size of surrounding watersheds to the cause of a flood in an underground bunker using data about the surrounding soil. The MHS curriculum emphasizes an understanding of water systems knowledge, as well as applying science knowledge in the argumentation tasks.

There were an additional 195 students who did not complete the MHS curriculum, and instead completed an online curriculum developed by Biological Science Curriculum Study (BSCS), that focused exclusively on water systems and did not include instruction aligned to argumentation. The BSCS curriculum Earth's Water Systems (EWS) was delivered online, and was designed to be a self-paced course that required little intervention from a teacher. EWS includes seven lessons intended for use in sequence. These seven lessons include: water cycle, surface water, groundwater, watersheds, atmospheric water, oceans, and human impact to water systems. Each lesson includes several informational slides with investigation sections that include a description of an experiment where the students are asked to make predictions and analyze the results. Each lesson concludes with a notebook-based exercise that asks students to apply their understanding to another context, often a more local context.

Data Collection

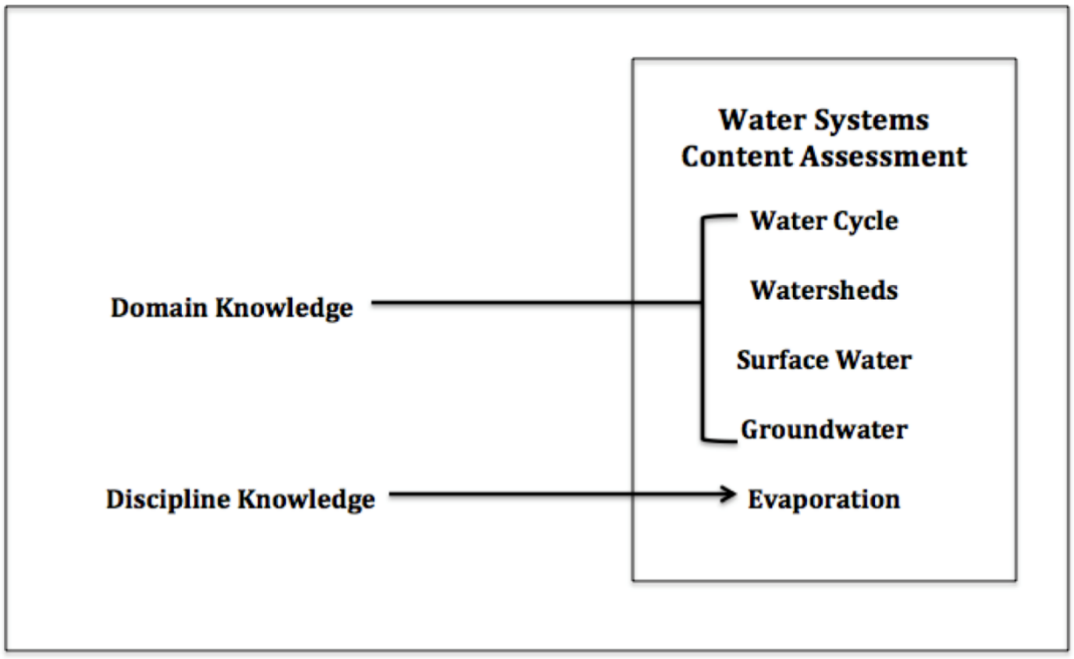
The purpose of this section is to provide an overview of the assessment instruments that were used within this study, as well as to outline the procedures for data collection that were employed. This section provides information about the specific assessments that were used within the study. These descriptions will provide details about previous pilot testing that had been conducted, and subsequent item revisions. Students who participated in this study completed both a water systems content assessment and an argumentation assessment. Both of these assessments were packaged together and delivered using Qualtrics, an online assessment platform. The section will conclude with a discussion about reliability, and describe the approach that this study takes towards estimating the internal consistency of both the water systems content and argumentation assessments.

Water systems content assessment

The water systems content assessment measures students' knowledge of water systems. Specifically, this knowledge includes a range of water systems topics including the water cycle, watersheds, surface water, groundwater, and evaporation. The main aim of this dissertation is to explore the connection between students' content knowledge and their ability to engage in argumentation. The argumentation assessment used to assess students' ability to engage in argumentation is grounded in the context of evaporation, thus the water systems assessment includes items that measure both domain knowledge and discipline knowledge relative to the argumentation assessment. Figure 3.1 below illustrates the ways in which domain knowledge and discipline knowledge map to the various topics covered on the water systems content assessment. Within the context of the

argumentation assessment, discipline knowledge refers to evaporation as the argumentation assessment scenario is involves students engaging in argumentation around an instance of evaporation. The remainder of the water systems content assessment includes domain knowledge since these are topics that are related to water systems but are not as specific to the argumentation assessment scenario.

Figure 3.1: Diagram of the content knowledge covered on the water systems content assessment.



Overview of assessment.

The water systems content assessment contains 24 multiple-choice questions. These questions cover a broad range of water systems topics including: groundwater, surface water, water pollution, evaporation, topography, and watersheds. Broadly speaking, the content of the assessment can be broken down into four areas. These include: groundwater, surface water, watersheds, water cycle, and evaporation. Several items required students to make sense of a diagram (e.g. watershed map, topographic map) in order to answer the questions. In this way, the vast majority of items on the

content assessment are application based questions in which students are required to apply their knowledge of water systems to answer a question or make a prediction, rather than simply recalling a specific term or definition. For example, many of the items that pertain to surface water present students with a watershed map with a river system clearly visible. Based on the map data, students are asked to predict which locations on the map would be impacted if water pollution occurred at a certain point on the map. In order to respond to these questions, students must be able to interpret a watershed map and be able to identify the direction of water movement to predict which sites would be impacted. The content assessment mostly contains these types of application questions, with each question including four possible answer choices. There are also a small number of items that ask for students to recall more factual information such as a description of evaporation or the source of energy for the water cycle. All items on this assessment were dichotomously scored (1 correct, 0 incorrect). Using the dichotomous scoring, sum scores were calculated for each response. These sum scores include on overall sum score for the entire test, as well as well as sub scores for each of the subcomponents of the assessment.

For the purposes of this study, the students' overall score on the entire water content assessment, student performance on the items that target knowledge of evaporation, and also student performance on all of the rest of the assessment items will be calculated. Students' knowledge of general water systems is framed in terms of domain knowledge as it encompasses a broad range of water systems topics that is consistent with the phrasing described by Alexander (1991), referring to general knowledge about a specialized field of study, i.e., water systems. When looking at the

field of water systems, Gunckel et al. (2012) developed a learning progression for water systems, and as part of that process, provides a detailed construct map for water systems. This construct map includes several environmental systems such as surface water, atmospheric water, and ground water. Each of these three topics are extensively covered in the items on the assessment, and looking at students' scores on the content assessment, after removing the evaporation items, will provide a measure of water systems domain knowledge that is consistent with the key ideas defined by the learning progression developed by Gunckel et al. (2012). Additionally, the discipline level knowledge is defined with respect to the argumentation assessment. Since the argumentation assessment is centered on a prompt that deals with evaporation specifically, the assessment taps into students' knowledge about evaporation when engaging in this assessment. This is consistent with the definition of discipline knowledge as specific knowledge about a specialized area of study (Alexander, 1991). In the Alexander (1991) framework for knowledge, discipline knowledge was described as knowledge about a more specific level of a broader domain. Evaporation can be conceptualized as a subcomponent of the broader water systems construct, as also described by the construct map positioned by Gunckel at al. (2012).

Summary of pilot test findings. The pilot testing of this assessment included data from 56 students ranging from 7^{th}-12^{th} grade. The results of this pilot test were used to inform revisions that would be made to the content assessment before use in a larger setting. Based on the pilot test, the assessment had fair internal consistency (alpha=0.54), but further analyses were used to better understand the item performance. The piloted version of the content assessment contained 31 multiple-choice items that related to key

concepts of water systems. A point-biserial correlation analysis was conducted on each item to better understand how students' performance on each item was related to their overall performance on the assessment. The results of the point biserial were used to inform the modification of items that may have been problematic and not indicative of student performance on the assessment as a whole.

Summary of revisions. Based on the correlation analysis, thirteen items demonstrated a weak point-biserial correlation (< .15), and those items were subjected to further review. Of those thirteen items, two were edited for clarity, and eleven were removed from the test, leaving twenty items. Another significant addition to the content assessment from the pilot testing was the addition of four items related to evaporation. This addition was made in an effort to increase the number of items that pertained to evaporation. Adding more items about evaporation was needed to allow for the exploration of the connection of more specific discipline knowledge and students' performance on the argumentation assessment. Items about evaporation were specifically added because the argumentation assessment is centered on an argumentation scenario that relates directly to evaporation. While the previous version of the content assessment did include two items about evaporation, this was not enough items to create a reliable sub-dimension that would be used in subsequent analyses. Based on the revisions, the water systems content assessment is a general water systems assessment that contains 24 items with six items that specifically target student understanding of evaporation.

Argumentation assessment

Overview of assessment. The argumentation assessment contains 12 multiple-choice questions. These questions cover a broad range of the argumentation construct as

described by the work of Sampson et al. (2014), and also includes items that fall at different levels of the Osborne et al. (2016) learning progression of argumentation. Broadly speaking, the assessment contains items that measure students' ability to identify critical components of an argument, align evidence to a given claim, and engage in critique. Within the assessment, there are four items relating to argument structure, four items that require students to align evidence to a claim, and four items that measure ability to engage in critique. The assessment is grounded in the context of evaporation, as students are presented with a scenario where a student leaves for a long summer vacation and leaves a bowl of water on his porch, and upon returning, finds the water is gone. Based on the observation that the water is gone, the assessment provides students with varying arguments about what could have happened to the water. This setting provides a context situated within water systems for students to engage in argumentation. All items on this assessment were scored in a dichotomous manner (1 correct, 0 incorrect).

Summary of pilot test findings. The pilot testing of this assessment included data from 56 students ranging from 7^{th}-12^{th} grade. The results of the pilot test helped inform the future design of the argumentation assessment. This version of the argumentation assessment included three different scenarios, which included dissolved materials, infiltration, and evaporation. With the assessment containing three scenarios, the instrument had an acceptable level of internal consistency (alpha=0.73). We also analyzed the reliability for each scenario individually, yielding the following: dissolving materials (alpha=0.51), infiltration (alpha=0.42), and evaporation (alpha=0.65). Reliability information was used to better understand the way that various scenario items were performing, and we also used to inform future iterations of the assessment.

Summary of revisions. The version of the argumentation assessment that was used in the pilot test was a much longer assessment as it included three distinct argumentation scenarios. Each of these three scenarios drew inspiration from the assessment items created and described by Osborne et al. (2016). These scenarios included the following topics: dissolving materials, evaporation, and infiltration. Based on the results of the pilot test, we decided to only include the evaporation scenario in future iterations, given the time constraints of the implementation period as well as considering the internal consistency of the three scenarios. Another major change from the pilot-test was to revise the critique items to be completely multiple-choice. During the pilot test, the critique items were a mixture of both open-ended and multiple-choice, and students had to write in their own words why they believed one argument was more convincing than another. While there are apparent benefits to the open-ended item design, given the time-constraints and the moderately strong correlation (0.425) between the multiple-choice and open-ended critique items, ultimately the open-ended items were removed to reduce the time required for students to complete the assessment. These data allowed us to remove the open-ended items, which decreased the time needed for students to complete the exam. While findings across all three scenarios were comparable, we decided to remove both the dissolving materials and infiltration prompts due to their length. Both of these scenarios were much longer than the evaporation scenario, so only using the evaporation scenario allowed the argumentation assessment to become shorter. Based on the revisions, the version of the argumentation assessment that will be used in this investigation contains 12 items with three sub-dimensions. These sub-dimensions include: alignment of claim and evidence (4 items), argumentation structure

(4 items), and argument critique (4 items). As noted above, these items all pertain to the evaporation scenario, which requires students to engage in argumentation centered on an issue that is grounded in the water systems discipline specific knowledge of evaporation.

Perspective of Reliability

Reliability is defined as the squared correlation between an observed score and a true score (Lord & Novick, 1968). Generally this is expressed as a ratio of the true score to the observed score. There are many ways to calculate reliability, but in classical test theory, reliability takes a singular value that is a description of the average error variance for all scores (Traub & Rowley, 1991). Commonly, reliability is reported using Cronbach alpha values. Cronbach (1951) described the general alpha value that is a special case of the Kuder-Richardson coefficient of equivalence that is shown to be the mean of all split-half coefficients resulting from various splittings of the test. The split-half method of measuring reliability is done by splitting a test in half and comparing the responses within each half (Callender & Osburn, 1977). The split-half method often yields a reliability value that can be inflated by a long test, and this is also true for Cronbach alpha (Dunn, Baguley, & Brunsden, 2014). Cronbach alpha is a singular value that represents the internal consistency of a test. In contrast to the simplistic nature of reliability within classical test theory, reliability in Item Response Theory (IRT) can be described as a function of proficiency or ability (Θ), and therefore, precision in IRT is not conceptualized as a singular value. Measurement precision in IRT is usually described in terms of $I(\Theta)$, the information function, the conditional error variance, or the standard error, and all of these are functions that vary with Θ. Although measurement error may vary as a function of proficiency, Green, Bock, Humphreys, Linn, and Reckase (1984)

note that it can also be averaged to give a marginal reliability that is more comparable to that of classical test theory. Essentially marginal reliably is an average of the information function across a range of all theta values, but in the computation of the marginal reliability, there is some loss of information in adding the error variance terms (Green et al., 1984).

For the purposes of this study, reliability will be conceptualized based on the IRT framework. As such, the perspective on reliability taken within this study will view measurement consistency as a function of a student's level of ability, rather than a singular value across the entire assessment. In this way, measurement error will be given a finer grain of detail that would be lost in reporting an overall reliability value. In analyzing the reliability of both the water systems content and argumentation assessment, information functions will be reported for each of the assessments. In addition, information functions will also be generated for each item on both the content assessment and the argumentation assessment. The item specific information functions will not be reported in this study. Within this analysis, a marginal reliability will also be computed for both the content and argumentation assessments, as well as the various sub-dimensions that these instruments contain. Reporting both the marginal reliability and the information function allows for both a overall picture of reliability across all student ability, but also allows for a more descriptive account of how measurement error changes depending on the ability level of students.

Data Analysis

This section of the chapter outlines the methodologies employed for data analysis. For coherence, this section has been organized around the three research questions of this

study. For each research question, descriptions of the techniques used are provided, as well as justification for use of these methods.

RQ1: What is the relationship between student's level of content knowledge and their ability to engage in scientific argumentation?

Structural equation modeling (SEM) is a general statistical modeling technique that is widely used within educational research. SEM has been described as a combination of factor analysis and regression or path analysis (Hox & Bechger, 1998). SEM provides a general framework for statistical analysis that includes several more traditional multivariate techniques. Often, SEM is visualized as a path diagram that graphically shows the connection between the variables of interest. SEM is a broad class of analyses, which can include both observed measured variables, as well as latent unobserved variables. SEM is similar to traditional methods like correlation and regression (Kline, 1998). Regression and SEM are both based on linear statistical models, and both share many of the same assumptions including normality, and neither technique offers a test of causality.

Despite the similarities between regression and SEM, there are differences between the two techniques that make SEM a more appropriate choice to use to answer this research questions. First is the issue of multicollinearity for multiple regression models. Essentially this means that when using a multiple regression, it is problematic for strong correlations to exist between the model predictors. Since this research question is interested in analyzing the connections between different levels of content knowledge (domain knowledge and discipline knowledge), it is likely that these will be tightly correlated. This correlation would be problematic for a multiple regression analysis, but

not for SEM. Additionally, SEM does not make the assumption that all data are collected without error, while a multiple regression assumes that measurement is without error (Hu, Bentler, 1999). Finally, since SEM has been used extensively as a confirmatory technique, and thus it will allow for evaluation of the proposed model of the connections between content knowledge and argumentation.

Figure 3.2: Proposed model of the relationship between content knowledge and argumentation

The proposed model of the relationship between content knowledge and argumentation includes three proposed latent parameters. These parameters include: argumentation, domain general content knowledge and discipline specific content knowledge. Argumentation in the model refers to the ability for students to engage in argumentation. The latent trait for students argumentation ability is defined based on the 12 items on the argumentation assessment. The other two latent constructs in the model are domain level content knowledge and discipline specific content knowledge. Data for these latent

variables will be obtained from the water systems content assessment. The definitions of the latent variables in the proposed model are expectations based on the conceptual nature of the assessment items themselves, but in order to determine the SEM structural model an exploratory factor analysis will be conducted on both the argumentation and water systems assessments to determine the number of latent variables in the model. Item will be linked to these latent variables in accordance with the factor loadings obtained from the exploratory factor analysis. This work will utilize principal axis factoring (Pituch & Stevens, 2016) and examine the communalities of the loadings of each item to determine the amount of variance in these items variables that can be accounted for by the factors. Additionally, there is a connection between both types of content knowledge and argumentation. The use of SEM will help provide evidence to confirm this relationship as well as provide estimates for the nature of these effects of the connections.

SEM will be used to explore this research question and evaluate the fit of the proposed model to the assessment data. Furthermore, since SEM is a conformity technique, it allows for the designation of the leadings of the observed variables (assessment items) to the latent traits a priori, specifically with respect to the discipline specific and domain level content knowledge. In this way, items that pertain to the evaporation will be considered discipline specific, and in the model load onto the discipline specific latent trait. This ensures that items that are indeed considered discipline specific do in fact relate to the specific content knowledge that is leveraged in the argumentation assessment.

In answering this research question, all analyses will be conducted using the open source R program, with use of the add-on packages lavaan, psych, and rio. These

analyses will include testing the assumptions of SEM, namely multivariate normality, and complete data. R will be used to also generate the path models with the use of the lavaan package. Output from R will produce a SEM diagram that includes the path coefficients between each of the observed variables in the model. These coefficients represent standardized versions of the linear regression weights that can be used to estimate the effect of content knowledge on argumentation. In addition to generating the path models, R will also be used to check for possible model misspecification using an analysis of the residual connivance matrix. Finally, the fit of SEM model will be evaluated. The analysis of model fit will be centered on two commonly used fit statistics within the literature; these include the root mean square error approximation (RMSEA) and confirmatory fit index (CFI). Both of these fit statistics will be calculated using the lavaan package within R. Fit will be assessed using widely referenced criteria for RMSEA less than 0.05 and a CFI of greater than 0.9 (Rigdon, 1996), which would both indicate an acceptable level of model fit.

RQ2: How do the different dimensions of argumentation vary in difficulty?

The main purpose of this research question is to determine if the items on the argumentation assessment have difficulties that are consistent with the Osborne et al. (2016) learning progression for scientific argumentation. In the learning progression, Osborne et al. (2016) leverage cognitive load theory, which argues that certain dimensions of argumentation vary in difficulty due to cognitive demands. For example, engaging in critique is a cognitively more demanding task that generating an argument since it requires students to manage more cognitive linkages (or degrees of coordination) between various argument components. Based on the considerations of the cognitive

demands of the argumentation dimensions, we would expect items involving critique to be the most difficult. The argumentation assessment contains items that fall into three sub-dimensions, which include critique, aligning claims and evidence, and identifying argument components. Based on the Osborne et al. (2016) learning progression, we would expect items involving the identification of argument components to be the easiest for students. These items only require students to identify these pieces, and not make any judgments about how well the claim, evidence, and reasoning in these arguments support each other. As such, these items are simply asking students to pick out these components in a complete argument. Next, items that ask students to align a piece of evidence with a given claim should be more difficult than simply identifying a piece of an argument. Completing these items require students to provide a judgment about how well each claim supports the evidence listed. Finally, we would expect the most difficult items on the assessment to be the ones that involve critique. Since these items require students to not only make a judgment about which of two given arguments is better, it also includes items that ask for a justification for their decision. These items are the most difficult because they require students to first determine which argument is better, and then to focus their thought about why one argument is better than another one. Osborne et al. (2016) referenced cognitive load theory when making the case for the increased difficulty of critique as being partially attributed to the increased number of connections between claim, reasoning, and evidence that are required to navigate this task. Similarly, Ford (2008) also provided support for the notion that engaging in critique is more cognitively demanding than argument construction.

In order to determine if the items in the argumentation assessment follow the expectations above, this study will use item response theory (IRT). IRT models stipulate a nonlinear monotonic function to account for the relation between examinee level on a latent variable and the probability of a specific item response (Lord, 1984). The basic assumptions of IRT are that the item responses are unidimensional and locally independent. Unidimensionality implies that the set of items assesses a single underlying trait dimension, and local independence means that if ability is held constant, the test items are pairwise uncorrelated. Modeling in IRT contains several variables that can be accounted for. These variables include discrimination parameters, guessing parameters, and a parameter for the upper asymptote. There are various IRT models that contain a different combination of these parameters. For this purposes of this study, a 3 parameters logistic (3PL) model will be used. The 3PL model allows for the estimation of a discrimination parameter for each item, which provides a measure of how well a correct item response separates low ability from high ability. Additionally, the 3PL model includes a guessing parameter that will adjust the lower asymptote of the item function to account for the probability of guessing the correct answer. The inclusion of a guessing parameter will help provide better item fit for some items in the argumentation assessment since many of the items only have three choices. In such cases, the probability of guessing a correct answer is significant enough that it cannot be ignored. While much of the item analysis work done in science education fits a Rasch model, since this model does not account for guessing, this study will employ a 3PL IRT model. Additionally, it is likely that the 3PL will provide a better item fit, since the Rasch approach does not include a guessing parameter, and the Rasch model places the restriction that the item

discrimination is fixed at 1 for all items. In contrast, the 3PL model will estimate a discrimination parameter for all items and include a guessing parameter. Evaluation of item fit for the 3PL model will be assessed using both RMSEA and CFI fit statistics that are commonly used in IRT modeling (Reise, Widaman, & Pugh, 1993). Fit will be assessed using widely referenced criteria for RMSEA, less than 0.05 and a CFI of greater than 0.9 (Rigdon, 1996), which would both indicate an acceptable level of model fit.

The 3PL IRT model will be fitted to the argumentation assessment data to yield estimations of the difficulty parameters for each item. The IRT item difficulty parameters are standardized, so, for instance, if an item had a difficulty parameter of 0 that would mean that a student with an average level of argumentation ability would have a 50% chance of getting that item correct. Additionally, if the difficulty parameter was -2, which would indicate that a student with a level of argumentation ability that is two standard deviations below average would have a 50% chance of getting that item correct. The results of the IRT analysis will be reported in a table for each item, which will include the argumentation dimensions associated with that item, as well the estimated difficulty parameter from the IRT model. The table will be sorted with the easiest items at the bottom and the most difficult items at the top, which will allow for a comparison to the expected item difficulties to be made. In addition to the reporting of the IRT fit statistics and the item difficulties, the analysis will also include the information function as well as the marginal reliability of the assessment. These will allow for communication of the error of measurement as a means to understand the reliability of the argumentation assessment. All of the IRT analysis for this research question will be conducted in R using both the rio and the mirt packages.

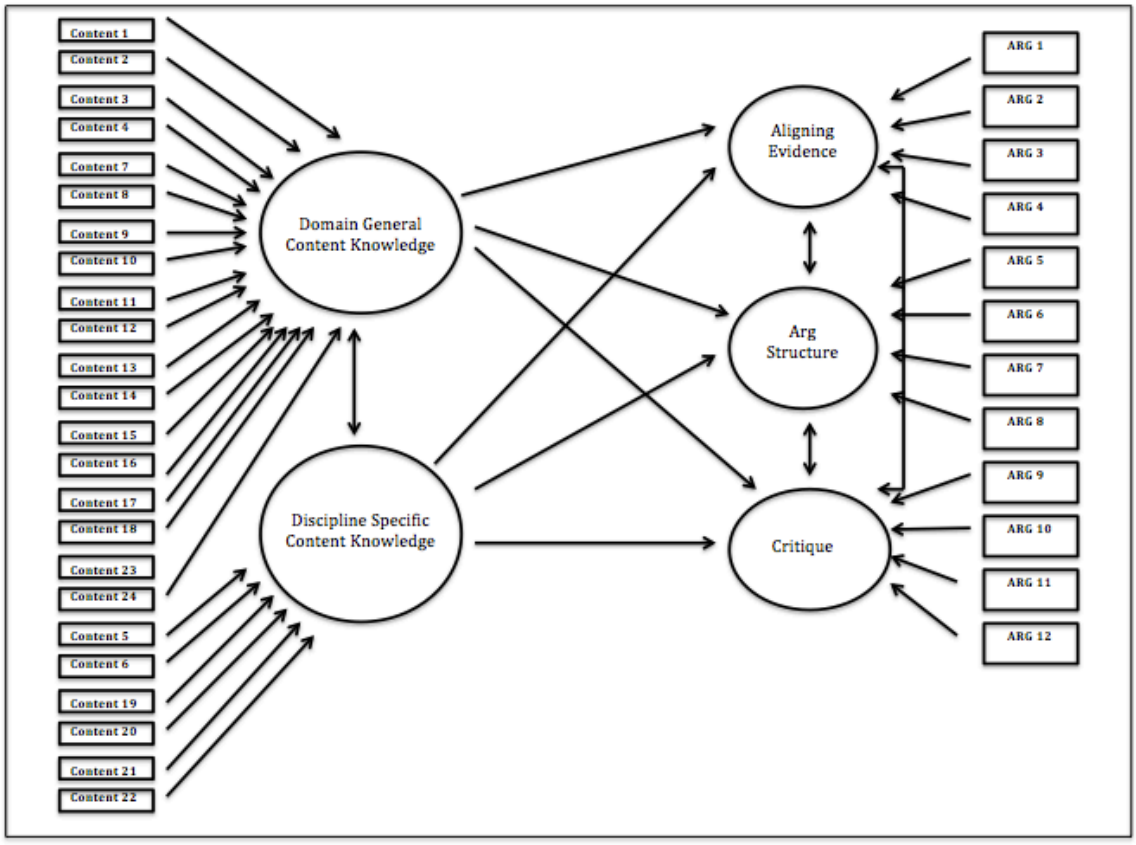
RQ3: How does content knowledge impact student performance on assessment items at different levels of an argumentation learning progression?

This research question builds on the results from the two previous questions. First, in order to investigate this question, it is assumed that the results of RQ2 suggest a separation of the argumentation sub-dimensions based on their difficulty level as estimated through an IRT model. The third research question essentially is an extension of the first research question, with a finer level of detail placed on argumentation. As such, the proposed model that will be evaluated in answering this research question will expand the argumentation construct to include the sub-dimensions of argumentation included in the assessment; these include argumentation structure, evidence alignment, and critique.

The proposed model for RQ3 includes all possible connections between both domain general and disciple specific content knowledge and the three dimensions of argumentation. It is important to point out that there are likely to be correlations between domain specific and discipline specific content knowledge. Additionally, it is also reasonable to expect correlations among the three dimensions of argumentation. In addition to the correlations, the proposed model also includes all possible connections between content knowledge and argumentation. While all possible connections are included in the model, it is expected that the magnitude of these connections will vary greatly. For example, it is expected that the discipline specific content knowledge will have a greater impact on argumentation ability than the domain general content knowledge. Furthermore, it is likely that the discipline specific content knowledge will also have a greater impact on the more difficult argumentation dimensions, namely

critique. Such an observed trend would be consistent with the work of Ford (2008) who describes a greater demand for content knowledge during the processes of critique.

Figure 3.3: Proposed Model for Research Question 3



In order to evaluate the proposed model, an Structural Equation Modeling (SEM) will be employed. SEM is used to evaluate the model in a similar manner to the way it was used in the first research question. This model contains the same latent variables domain general and discipline specific content knowledge, which will be defined in the same manner as in RQ1, based on the items from the water systems content assessment. The additional latent variables for the dimensions of argumentation will be assigned by using items on the argumentation assessment and will be defined a priori, similar to the latent variables on the water systems content assessment. The three latent variables for

argumentation are critique, aligning evidence, and argument structure. In order to define these latent variables, the results of the exploratory factor analyses (Pituch & Stevens, 2016) for both the water systems content and argumentation assessments will be used. The factor loadings will be used to assign which items load onto each of the latent traits. Once all of the variables are defined, R will be used along with the lavaan package to generate the path diagram. The proposed model will be modified based on the SEM results to only statistically significant path connections. Model fit will be assessed using widely referenced criteria for RMSEA, less than 0.05 and a CFI of greater than 0.9 (Rigdon, 1996), which would both indicate an acceptable level of model fit.

Chapter 4 Results

The purpose of this chapter is to report the findings of this study based on using the methodology that was described in the previous chapter. This chapter begins with an overview of the descriptive statistics for each assessment instrument, and also includes a discussion of the assumptions of IRT for both the content assessment and the augmentation test. These assumptions include unidimensionality, monotonicity, and local independence. Following the checking of these assumptions, this chapter will also include a discussion of reliability of the assessments by reporting the information function for each instrument. Following these analyses, the remainder of the chapter has been organized around the research questions of the study. The main aim of this research study is to examine the connection between students' level of content knowledge and their ability to engage in scientific argumentation. Sections of the chapter will report the results of the factor analysis on both of the instruments, results of the IRT analysis of both instruments, and discuss the structural equation modeling that was used to answer the research questions.

Data Cleaning

Initially, there were 917 students that responded to both the argumentation and content assessments. In cleaning the data, all instances of incomplete data were removed. After removing incomplete data, there were 834 remaining student responses. Next, in an effort to remove data from students that did not exert satisfactory effort on the test, test duration data was used to identify student that completed the assessments in a significantly short amount of time that it is impossible they took the time to read and consider each item carefully. Since these assessments were administered via qualtrics,

each response included a value of the number of seconds needed for the student to complete the entire assessment. Outliers in terms of test duration were removed by converting all values of test duration into a standardized z score, and all z-score values above 3 and below -3 were removed. This results in a final sample of 808 students that were used in subsequent analyses.

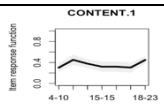
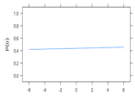
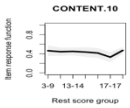
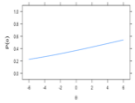
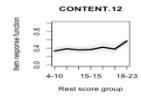
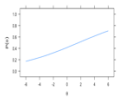
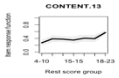
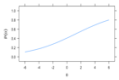
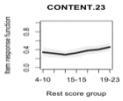
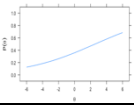
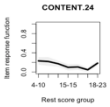
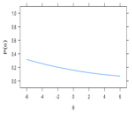
Summary of Removed Items

The decision to remove items from these analyses was informed by the output of the IRT model. After cleaning the data and fitting a 2PL model, which allows for an approximation of a discrimination parameter for each item, both the discrimination parameters and the rest scores, students' performances on the rest of the test were used to determine if an item should be removed. A rest score is similar to a point biserial analyses in that it compares the correlation between the student getting an item correct, and that student score on the rest of the assessment. A rest score plot evaluates the probability of a student with a certain score on the test getting a specific item correct and these probabilities are plotted as a function of the students content test score. Ideally, these functions should be increasing throughout, and students getting an item correct should correlate to a higher score on the test overall. Another parameter that was considered was the item trace plots. A trace plot shows a visual representation of how well an item discriminates students of low ability and high ability. Trace lines plot the probability of a student getting an item correct given their level of ability as estimated by the IRT model. Ideally these should look like a sigmoidal "S" curve, and the steepness of the curve is a representation of the discrimination parameter for the item. De Ayala (2013) recommends looking at both of these parameters when deciding to remove an item

from an analysis. Below is a summary of the items that have been removed from both the content assessment and the argumentation assessment based on this criteria.

Content Assessment. The content assessment consisted of 24 items in total. Based on the analysis of both the item discrimination parameters and the rest plots, six items have been removed from the assessment leaving 18 items. Figure 4.1 below includes a plot both the rest scores the trace lines for each item removed from the content test.

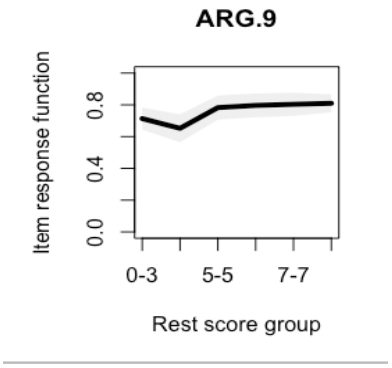
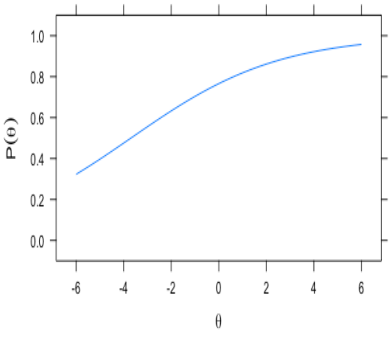
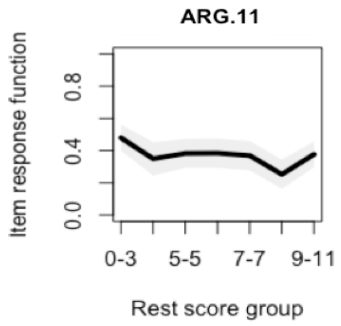
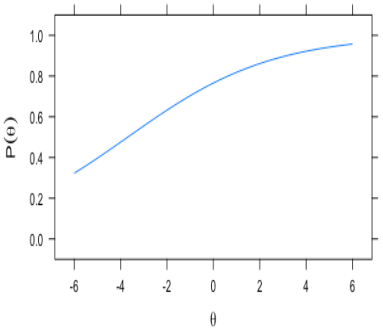
Figure 4.1: Summary of the items removed from the content assessment. This figure includes both rest plots and item trace lines.

Item #	Rest Score Plot	Plot of Item Trace Line
Item 1		
Item 10		
Item 12		
Item 13		
Item 23		
Item 24		

Argumentation Assessment. The argumentation assessment consisted of 12 items in total. Based on the analysis of both the item discrimination parameters and the rest plots, two items have been removed from the assessment leaving 10 items. It is worth pointing out here that both of the items removed were critique items, but there were

different from the items on the rest of the test in that they only contained two answer choices, and both items presented students with two competing arguments, and the student had to select which argument they felt was better. Figure 4.2 below includes a plot of both the rest scores and the trace lines for each item.

Figure 4.2: Summary of the items removed from the argumentation assessment. This figure includes both rest plots and item trace lines.

Item #	Rest Score Plot	Plot of Item Trace Line
9		
11		

Summary of Descriptive Statistics. This section provides a brief overview of standard descriptive statistics based on student scores on each of the assessments. Figure 4.3 below presents basic descriptive statistics for both the argumentation and content assessments. Due to the removal of items detailed above, the maximum possible score on the argumentation assessment was 10, and the maximum possible score on the content

assessment was 18. When looking at the means for the two assessment based on percentage correct, the content assessment has a mean of around 13 out of 18 possible, and the argumentation assessment had a mean of 5.24 out of a possible 10. This means that on average students answered a higher number of items correct on the content assessment than the argumentation assessment, the issue of item difficulty will be explored in greater detail later when discussing the use of IRT analysis to approximate the difficulties of the items on the argumentation assessment.

Figure 4.3: Summary of descriptive statistics for the argumentation and content assessments

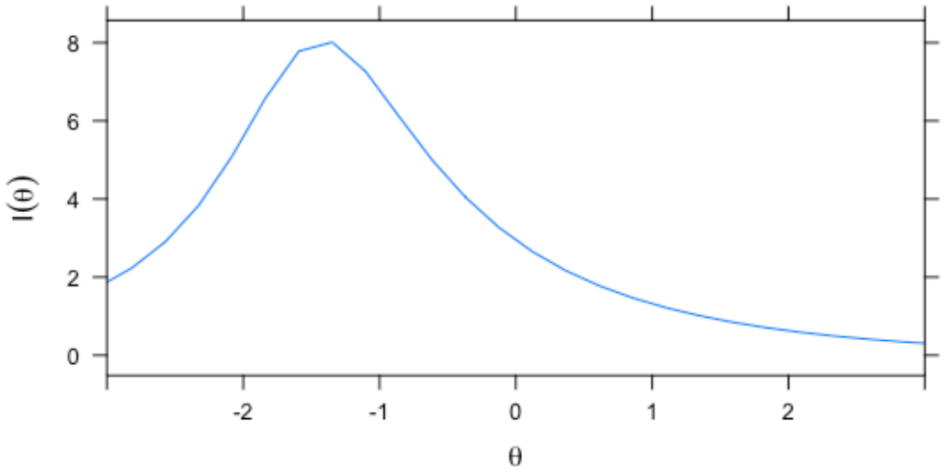
Instrument	Mean	St.Dev.	Median	Max	Min	n
Argumentation Assessment	5.24	2.34	5	10	0	808
Content Assessment	12.92	3.34	13	18	3	808

Summary of Reliability

When looking at the reliability, both instruments provide fairly reliable measures of both argumentation(α=.67) and water systems content knowledge (α=.74). According to the Pituch & Stevens (2016) an alpha coefficient above .7 is a generally an acceptable level of reliability. While the water systems content assessment meets this criteria, the argumentation assessment is approaching this desired value. While several researchers have raised concerns over use of an alpha statistic as the sole metric of reliability (Pituch & Stevens, 2016), work in the field of IRT often uses information as a means to show instrument reliability (De Ayala, 2016). Information is generally reported as a function that plots reliability as a function of student ability. As a result, information allows the

researcher to gain a more detailed view of reliability that can vary based on the ability level of the students. Figure 4.4 below plots the information function for the water systems content assessment.

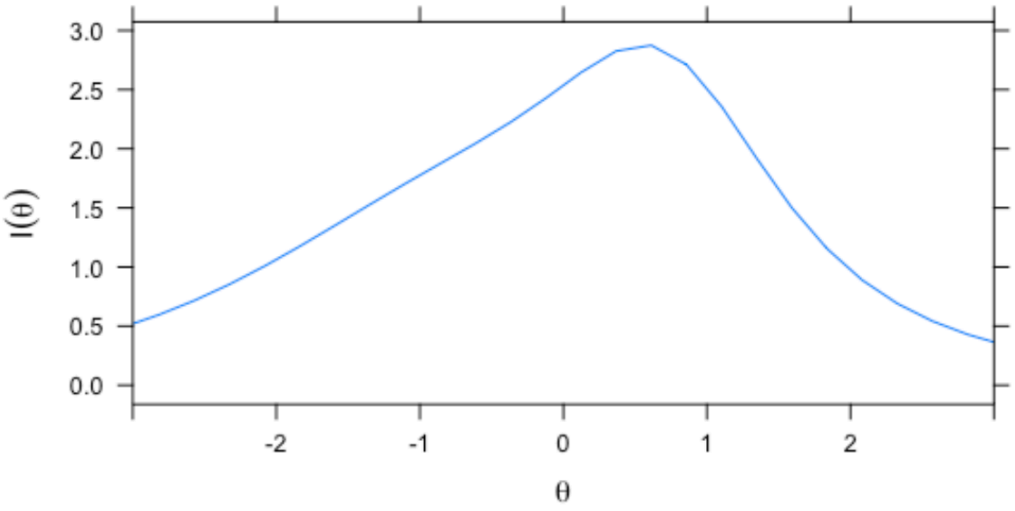
Figure 4.4: Information function of the Water Systems Content Assessment



Based on the information function, the water systems content assessment is a reliable measure of water systems knowledge for most students. Specifically, the information is highest for students of lower than average ability, but is also reliable for students up to two standards deviations above average water systems knowledge. Additionally, Figure 4.5 reports the information function for the argumentation assessment. Based on the information plot, the argumentation assessment provides a reliable measure of argumentation competency for a broad range of student ability ranging from three standard deviations above and below average argumentation competency. In addition to information functions, researchers using IRT often report a marginal reliability, which is generally a summative value that encompassed the entire information function. Marginal reliability is often interpreted the same as an alpha statistic from Classical Test Theory (De Ayala, 2016). When considering the marginal

reliabilities, both the water systems content (.72) and argumentation (.68) assessment have an acceptable level of reliability.

Figure 4.5: Information Function for the Argumentation Assessment



RQ1: SEM

The purpose of this research question is to investigate the relationship between students' argumentation ability and their level of content knowledge. This relationship was explored by analyzing student responses to both the content and argumentation assessments. In analyzing the student responses to both of these assessments, both exploratory factor analysis and structural equation modeling were used. Exploratory factor analysis was used to determine the number of factors of each assessment, and also to determine which items loaded on to these factors. The factor loading information was then used to describe and label the factors based on the types of items that loaded significantly (higher than .3) on them. The results of the factor analysis was used to create a hypothetical model of the relationship between argumentation and content knowledge that will be tested using structural equation modeling. This section of the

chapter will describe the result of the exploratory factor analyses conducted on both the argumentation and content assessments, as well as describe the use of SEM to investigate the relationship between students' performance on both the content assessment and the argumentation assessment.

Factor Analysis of Content Assessment

This section describes the means of assessing the extent to which the content assessment data meets the criteria and is suitable for exploratory factor analysis. Additionally, this section will also describe the results of the factor analysis with specific reporting of the scree plots for factor identification and the specific variable factor loadings that will be used in the subsequent SEM analysis. The results of the factor analysis will also report the fit statistics (i.e. RMSR, RMSEA) for the factor analysis model.

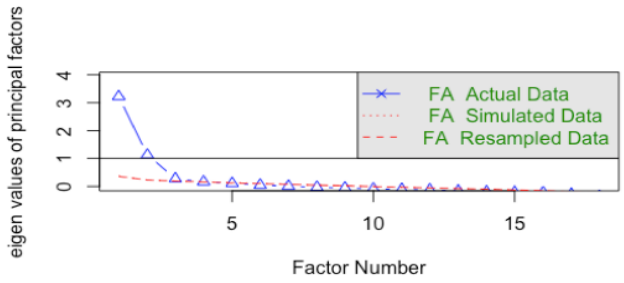
Assumptions for Factor Analysis. In order for data to be suitable for factor analysis, there are criteria that must be met. Predominately there are four main assumptions of factor analysis, which includes multivariate normality, linear relations between variables, factorability, and adequate sample size (Pituch & Stevens, 2016). While statistical inference is improved if the variables are multivariable normal (O'Rourke, Psych, & Hatcher, 2013) an assumption of multivariate normality is not made when using principle axis factoring in an exploratory factor analysis (Pitch & Stevens, 2016). Linear relationships between variables and factorability can be assessed using the Kaiser-Meyer-Olkin (KMO) test as well as Bartlett's test for sphericity (Pitch & Stevens, 2016). Factorability refers to the assumption that there are at least some correlations amongst the variables such that coherent factors can be identified. Essentially, there

should be some degree of collinearly among the variables but not an extreme degree of singularity among the variables such that they represent a single factor. In order for the data to satisfy these assumptions, the KMO should have a value greater than .6, and Bartlett's test should have a significant result. The content assessment data satisfies both the assumption of linear relationships between variables, and factorability (KMO= .84; $X^2(153)=2491$, $p<.001$). Finally, the assumption of adequate sampling ensures that the sample size is large enough to yield reliable estimates of the correlations between variables. Recommendations for sample size for factor analysis typically use the N/k formula, where N is the sample size, and k is the number of variables. Pituch & Stevens (2016) recommend at least a 20:1 ratio between the sample size and the number of variables. For this analysis the sample size is 808, and there are 18 items, which yields a ratio of 44.6:1, well above the recommended 20:1. Based on these criteria, the content assessment data is suitable for exploratory factor analysis.

Results of Factor Analysis. The section describes the results of the exploratory factor analysis on the content assessment data. In order to determine the number of factors to include in the factor model a scree test (Cattrel, 1988) was conducted. A scree test is a graphical method that plots the magnitude of the eigenvalues on the vertical axis against their ordinal numbers (i.e. first, second, etc.). Generally, the magnitude of successive eigenvalues drops off significantly after the first couple of eigenvalues, and reaches a point where the successive values level off. Cattrel (1988) makes the recommendation to retain all eigenvalues (and therefore factors) in the sharp decent before the first one on the line where they level off. The results of the scree test are

shown below in figure 4.6. Based on the results of the scree test, and the recommendations two factors were retained and used in the factor model.

Figure 4.6: Scree Test results for content assessment



To generate the factor model, principal axis factoring was used as it is recommended for an exploratory analysis (Pituch & Stevens, 2016), and obliging rotation was used since there was expected some level of correlation between the two factors since they both relate to students performance on the content assessment. Figure 4.7 provides an overview of the factor loadings for each of the two factors retained in the model. With the use of oblimin rotation simple structure was achieved, as there are no items that load significantly on both factors. It should also be noted that Figure 4.7 reports the factor loadings as well as the communalities of the items (h2). Communality refers to the extent to which an item correlates with all other items. Higher communalities are better generally if an item has a communality below .4, then it may struggle to load significantly to any factor (Pituch & Stevens, 2016). The results show that 5 items load to factor 1, and 12 items load onto factor 2. Also, assessment items #2 did not load significantly to either factor, and for that reason it will not be included in the subsequent SEM analysis. Based on the loadings, factor 1 seems to be related to the topic of evaporation, as all of the items that load to it pertain to the topic of evaporation. The items loading to factor 2 cover a wide range of water systems topics including

groundwater, surface water, and watersheds. For this reason, factor 2 can be conceptualized as more a general factor about water systems content that is not specific to evaporation (as these items load to factor 1). It is important to note that there is one item about evaporation that did not load significantly to factor 1, and also that the items pertaining to the water cycle loaded rather strongly to factor 2 despite being closely related to the concept of evaporation.

Figure 4.7: Summary of factor loadings from the content assessment EFA (n=808, X2(153)= 2491, P<.001, RMSR= .04, RMSEA= .044

Item#	Topic	Factor 1	Factor 2	h2
2	Groundwater	-.03	.19	.031
3	Groundwater	-.11	.329	.291
4	Water Cycle	-.02	.557	.302
5	Evaporation	.671	.01	.454
6	Evaporation	.00	.458	.49
7	Watersheds	-.03	.333	.23
8	Watersheds	.02	.371	.37
9	Watersheds	.06	.407	.166
11	Groundwater	.09	.363	.164
14	Surface Water	.07	.362	.21
15	Surface Water	-.07	.445	.378
16	Surface Water	.17	.388	.23
17	Watersheds	.11	.369	.198
18	Watersheds	.14	.344	.175
19	Evaporation	.603	.14	.454
20	Evaporation	.792	-.04	.602
21	Evaporation	.673	-.1	.409
22	Evaporation	.644	.1	.477

The adequacy of model fit for the factor model was evaluated using both the Tucker-Lewis index and the RMSEA. Based on both of these metrics, there was an acceptable level of fit (Tucker-Lewis= .899; RMSR=.04). A generally acceptable value for the Tucker-Lewis index is .9 (Pituch & Stevens, 2016) and .899 closely approaches this value. The root-mean square error of the residuals should be less than .05 (Kline, 2015; Pituch & Stevens, 2016), and the value .04 satisfies this requirement. The results of this factor analysis will be used in the subsequent SEM analysis.

Factor Analysis of Argumentation Assessment

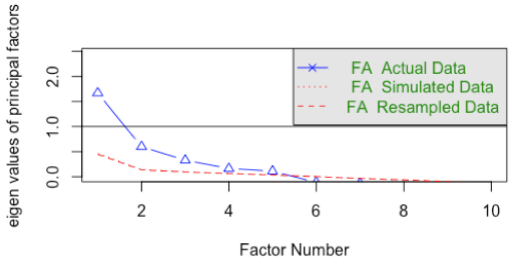
This section describes the means of assessing the extent to which the argumentation assessment data meets the criteria and is suitable for exploratory factor analysis. Additionally, this section will also describe the results of the factor analysis with specific reporting of the scree plots for factor identification and the specific variable factor loadings that will be used in the subsequent SEM analysis. The results of the factor analysis will also report the fit statistics (Tucker-Lewis Index, RMSEA) for the factor analysis model.

Assumptions of Factor Analysis. As indicated above in the section describing the content assessment there are four main assumptions of factor analysis, which includes multivariate normality, linear relations between variables, factorability, and adequate sample size (Pituch & Stevens, 2016). While statistical inference is improved if the variables are multivariable normal (O'Rourke, Psych, & Hatcher, 2013) an assumption of multivariate normality is not made when using principle axis factoring in an exploratory factor analysis (Pitch & Stevens, 2016). The argumentation assessment data satisfies both the assumption of linear relationships between variables, and factorability (KMO= .66;

$X^2(45)=1069$, $p<.001$). Additionally, the ratio of the sample size to the number of variables is 80.8:1, which is above the recommended 20:1 ratio. Based on these criteria, the argumentation assessment data is suitable for exploratory factor analysis.

Results of Factor Analysis. The section describes the results of the exploratory factor analysis on the argumentation assessment data. In order to determine the number of factors to include in the factor model a scree test (Cattrel, 1988) was conducted. The results of the scree test for the argumentation assessment are shown below in figure 4.8, which is a scree plot that shows the eigenvalues on the vertical axis, and the ordinal designation of these eigenvalues on the horizontal axis. The same criteria from the content assessment were used for selecting the number of factors to retain. This included all of the eigenvalues before the plot leveled off. Based on the results of the scree test, one factor will be retained and the analysis will describe a one-factor model for the argumentation assessment.

Figure 4.8: Results of Scree test on argumentation assessment



To generate the factor model, principal axis factoring was used as it is recommended for an exploratory analysis (Pituch & Stevens, 2016). Since the resulting model only includes a single factor, the analysis was used to determine if all ten items from the argumentation assessment actually load onto the singular factor at a significant level (greater than .3). Figure 4.9 below provides the factor loadings for all ten items

onto the one factor. Based on the loadings, we find that eight of the ten items load significantly (greater than .3) to the one factor, which can be described as broadly as competency in scientific argumentation. There is one item that pertains to aligning evidence to claim, and another item that relates to identifying argumentation structure that do not load significantly to the one factor of scientific argumentation. Based on these results, eight of the ten items will be used in the subsequent SEM analysis, with all eight items loading onto one factor.

Figure 4.9: Factor loadings for one factor model of the argumentation assessment data. X2(45)=1069, p<.001, RMSR= .1 RMSEA= .016.

Item #	Topic	Factor 1	h2
1	Align Evidence	.45	.44
2	Align Evidence	.49	.42
3	Align Evidence	.20	.09
4	Align Evidence	.37	.21
5	Structure	.41	.38
6	Structure	.45	.37
7	Structure	.26	.17
8	Structure	.30	.18
10	Critique	.60	.49
12	Critique	.40	.42

Using Structural Equation Modeling

This section describes the use of structural equation modeling to investigate the relationship between students' level of content knowledge and their ability to engage in scientific argumentation. This section will begin with an overview of the proposed model of the relationship that will be used using SEM methods. The proposed model was

informed by the results of the factor analysis that have been conducted previously in this chapter on both the content assessment and argumentation assessment.

Before explaining the model building and evaluation process, it is important to discuss the methods of model estimation and fit evaluation that will be used in this analysis. The most common method of estimation that is used in SEM is the maximum likelihood estimator. As noted by Kline (2015) the term maximum likelihood (ML) described the idea that the estimates are ones that maximize the likelihood given the observations. ML estimation is a normal theory method that makes the assumption of multivariate normality of the endogenous variables. Since only continuous variables can have normal distributions, and therefore the dichotomous test data that was collected in this study does not meet the requirements for ML estimation. Kline (2015) suggests using alternative estimation techniques when working with categorical or binary data. Muthen, de Toit, and Spisic (1997) describe the robust weighted least squares (RWLS) estimation, which uses simpler matrix calculations than the full WLS method that does not make the assumption that the observed variables in the SEM are normally distributed. As such, the RWLS method of estimation is often used when dealing with categorical or binary data (Kline, 2015; Muthen, du Toit, & Spisic, 1997). Finney & DiStefano (2013) report that the RWLS estimation performs well in computer simulation studies when the sample size is larger than 250. As such the RWLS method is appropriate for this analysis as the sample is sufficiently large (N=808) and the data are not normally distributed.

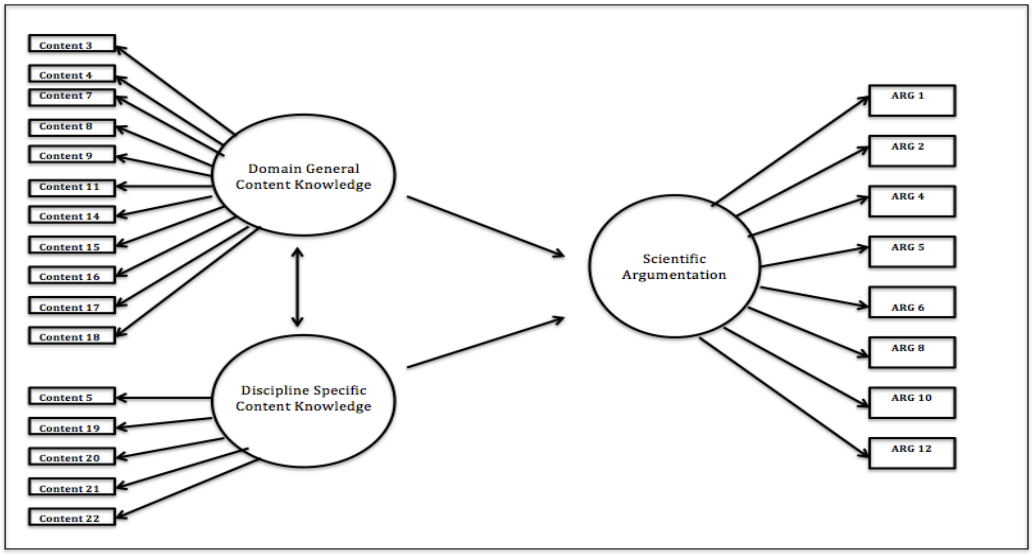
In addition to estimation, the analysis of model fit will include use of the chi-square test, Steiger-Lind Root Mean Square Error Approximation (RMSEA; Steiger, 1990), Bentler Comparative Fit Index (CFI: Bentler, 1990), and the Standardized Root

Mean Square Residual (SRMR). The Chi Square test is the most basic model test statistic (Kline, 2015) and it tests whether the population covariance values are consistent with those predicted by the model. As such, a model that fit perfectly to the data would have a X^2 value of zero and a statistically non-significant p value above .05. The CFI is an incremental fix index that measures the relative fit of the target model against an independent model (where all covariances are equal to zero) The RMSEA is based on the X^2 value, but is adjusted for model parsimony (degrees of freedom) and sample size. The SRMR is a measure of the average absolute correlational residual (i.e. difference between observed and predicted correlations). Kline (2015) described the following criteria for the desired values of each of these fit indices: X^2 value should be close to zero, and have a p >.05, CFI should be greater than .95, RMSEA should be below .08, and SRMR should be below .08. These recommendations for fit statistics will be used in this analysis of the overall fit of the structural models.

Model Specification. Based on the results of the factor analysis, this section will provide an overview of the proposed SEM model that will be tested with further analysis. Figure 8 below is a diagram of the SEM model that will be tested in this analysis with the identification of the latent traits based on the outcomes of the previous exploratory factor analyses. The model breaks down the content assessment into two factors (one a general water systems factor loaded by item from various topics) and a second evaporation factor that is loaded on by five items that cover evaporation. The argumentation assessment contains one factor that is loaded on by eight items covering all tested dimensions of argumentation (i.e. critique, structure, aligning evidence). SEM will be used to assess the fit of this model, as well as estimate the connections between the latent constructs.

Model Modification. The proposed model needed to be modified due to the failure of the model to converge. While failure of the model to converge is not ideal, it can happen in several circumstances. The model estimation strategies that are used in SEM are iterative in nature, meaning that the computer will derive an initial solution and improve upon the solution through additional cycles (Kline, 2015). With this technique the model continue to improve

Figure 4.10: Proposed SEM model of the relationship between content knowledge and argumentation



over successive iterations of the estimation. The model reaches convergence when the iterations no longer improve the overall fit of the model, and that point the model coefficients can be estimated and reported. Failure to converge occurs when the computer fails to find an appropriate solution. Kline (2015) makes the suggestion of "building your model backup" when dealing with non-convergence and look to simplify the model wherever possible. Non-convergence can occur for over specified models, meaning that there are too many connections between latent constructs. Kline (2015) also notes that a common cause of non-convergence are poor starting values for the model, and makes the

recommendation to try to improve these values by correcting initial coefficient estimates that are expected to be far off. Based on those recommendations, the first step was to evaluate the starting values for the model and determine if these could be improved by identifying any values that were obviously incorrect. An initial inspection of the model coefficients did not result in any changes made to the starting values, as all initial estimates seemed appropriate. In some structural models the researcher may want to fix the coefficients to certain values based on the previous literature, but for this analysis all model coefficients were left free and this estimated by the model as there was no literature estimate to estimate these parameters since there is no published work including the argumentation and content assessments to base the model coefficients on.

Following the analysis of the model starting values, the next step was to attempt to simplify the model to allow for convergence. In order to improve the model, the statistical significance of each model coefficient was analyzed. Based on this analysis, six items in the content assessment and four items in the argumentation assessment were not statistically significant (all $p > .05$). As a result, these six content item (2,7,8,14,17,18) and four argumentation items (3,4,7,8) were removed from the model. It should also be noted that the items removed has lower communality values in the factor analysis indicating that their factor loadings alone did account for a significant amount of the variance observed in these variables. After this modification, the resulting updated model was run and the model converged. While the updated model was able to converge, there were concerns around the fit of the model. As mentioned earlier, the main fit statistics that were used in this analysis were the chi square test, RMSEA, RMSR, and CFI. While the updated model did have fit statistics that approached acceptable values ($X^2(153)=$

761, p<.001; RMSEA=.075; RMSR=.089; CFI=.878), additional model was needed to improve the fit to reach commonly accepted values.

Further model modification was conducted using the modification indices. Modification indices are calculated by the computer and include statistically significant model connections that currently do not exist on the model. Kline (2015) advices care in looking at the modification indices, as to not instantly include all significant connections into the model, which can result in over fitting. As such, the only modification that were made to the model was the addition of covariance lines between the items that load to the various latent traits. This resulted in covariance lines being added to the model between the items on the argumentation assessment, evaporation items on the content assessment, and general water systems items on the content assessment. This resulting model was fitted, and evaluated in the next section of this chapter.

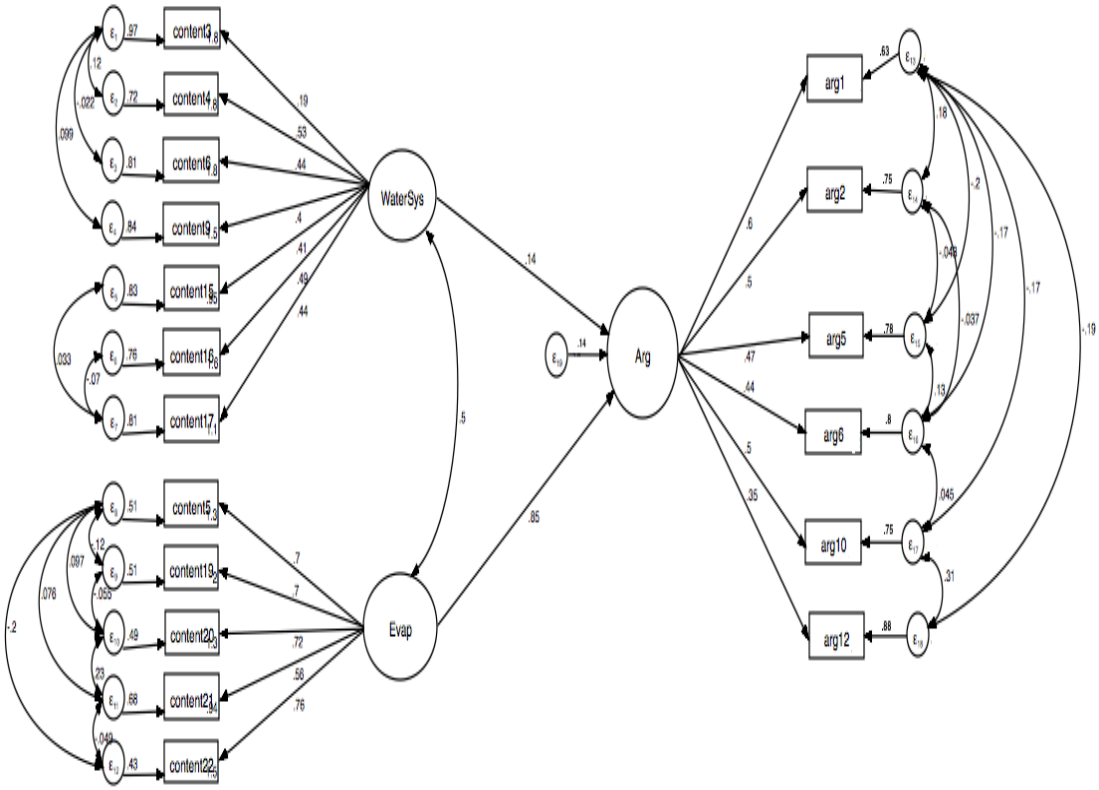
Results of SEM. All participant responses were subjected to SEM analysis (n=808). Due to the variables being dichotomous, SEM was estimated using the robust weighted least squares method (Muthen, du Toit, & Spisic, 1997). Correlational residuals were examined and all values were below .10, which has been suggested as a cutoff to where correlation residuals higher than .10 require special attention as this can signify poor local fit (Kline, 2015). Figure 4.11 reports the fit indices of the structural model, as well as the recommended values reported within the literature.

Figure 4.11: Fit Indices for the Final Structural Model (n=808)

Fit Index	Value	Recommended Value
Chi-square	264	-
Chi-square per degree of freedom	2.37	<5
Root mean square error approximation (RMSEA)	0.042	<.08
Standardized Root Mean Square Residual (SRMR)	0.038	<.08
Comparative Fit Index (CFI)	0.95	<.95

Included in Figure 4.11 is an adjusted chi-square value that is useful with samples larger than ~400, which divides the chi-square statistic by the degrees of freedom and this value should be less than 5 (Lee, Johanson, & Tsai, 2008). The fit statistics in Figure 4.11 show that all of the reported statistics are within the recommendations of the literature, indicating there is an acceptable amount of model fit of the structural model that was estimated using SEM. The structural model from the SEM analysis is reported in Figure 4.12. The structural model shows the estimation for the path coefficients between the variables included in the model. The coefficients included in Figure 4.12 are standardized, and are interpreted as a standardized regression coefficient.

Figure 4.12: Final structural model of the relationship between content knowledge and scientific argumentation.



Within the SEM, the main two coefficients of interest are the connections between the water systems latent trait and argumentation, as well as the connection between the evaporation latent variable and argumentation. The estimated coefficient between evaporation and argumentation is .85 meaning that an increase of one standard deviation of evaporation variable is associated with an increase of .85 standard deviations in the argumentation latent variable. This is considerably higher than the coefficient linking the water systems latent variable with argumentation, which is .14. It is important to note the non-statistically significant relationships are not depicted in the model.

RQ2:IRT Analysis of Argumentation Assessment

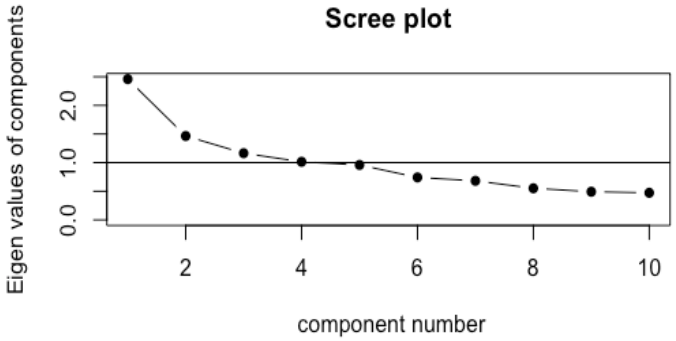
The purpose of this research question is to explore the extent to which the assessment items on the argumentation assessment corresponds to various levels of an argumentation learning progression. These dimensions of argumentation include: alignment of claim and evidence, argument structure, and critique. The goal of this analysis is to fit an IRT model to the argumentation assessment data, and approximate the item difficulty parameters and compare them across items that correspond to various dimensions of argumentation.

IRT Assumptions

The first step in this analysis is to evaluate the appropriateness of using IRT by testing that this data set meets the assumptions of IRT. IRT has three main assumptions: unidimensionality, monotonicity, and local independence. In order to assess unidimensionality, eigenvalues were calculated and plotted in a scree plot shown in Figure 4.13. While the scree plot approximately shows unidimensionality, there is some ambiguity in determining the number of components. In addition to using a scree plot to

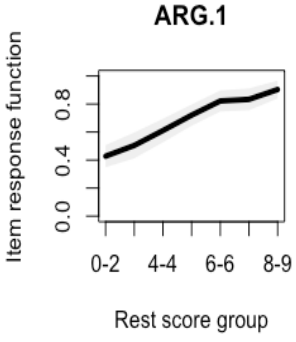
determine unidimensionality, De Ayala (2013) also describes the process of comparing the ratio of the first eigenvalue to the sum of the total. If the value of this ratio is greater than .2, the data meets the IRT unidimensionality requirement. In this case this value is .24, meaning that the argumentation assessment data satisfies the IRT requirement of unidimensionality.

Figure 4.13: Scree plot of the Eigen values of components from the argumentation assessment



The second assumption of IRT is mononicity, which is generally determined by analyzing the rest score plots for each assessment item. Checking for monotonicity insures that all test items are correlated with better overall student performance on the assessment. For this reason, an ideal rest score plot should be increasing as the students rest score increases. All of the item rest score plots satisfied the assumption of monotonicity, and Figure 4.14 presents a sample rest score plot that is representative of the other assessment items.

Figure 4.14: Sample rest score plot for the argumentation assessment.



The third assumption of IRT is local independence. Essentially, local independence is the principal that all observed items are conditionally independent of each other given an individual score on a latent trait. The means that the latent trait accounts for the relationships that are observed between items, and thus an incorrect response to one item should not mean an incorrect response on another item. In order to assess the local independence of the assessment items, a Q3 statistic was used that was based on the residual for each student's response to each item. After computing the residuals, the Q3 statistic was computed as the linear correlation between the residuals for each item. This generates a table that is a correlation matrix showing the correlations of the residuals between all items on the assessment. Houts & Cai (2013) suggests that a LD X^2 value above 3 or below -3 cannot be ignored and should be further reinvestigated. Upon investigation, all of the LD X^2 values were between 3 and -3, which satisfy the assumption of local independence of all assessment items.

IRT Model. After checking to ensure the argumentation assessment data meets the assumptions of IRT, a model was fitted to the data. While it was the intention to fit a 3 parameter logistic (3PL) IRT model, ultimately a 2 parameter logistic model (2PL) was used. This decision was made based on the fact that the 3PL model converged after 1110 iterations, and produced nonsensical estimates for the various item discrimination parameters. Additionally, when fitting the 3PL model many items had the estimation of the guessing parameter at 0, suggesting that the 3PL model is not useful for interpreting this data. As a result, a 2PL model used, which converged after 30 iterations. The 2PL model estimates an item discrimination parameter and a difficulty parameter for each of the ten items on the argumentation assessment. The coefficients from the 2PL model are

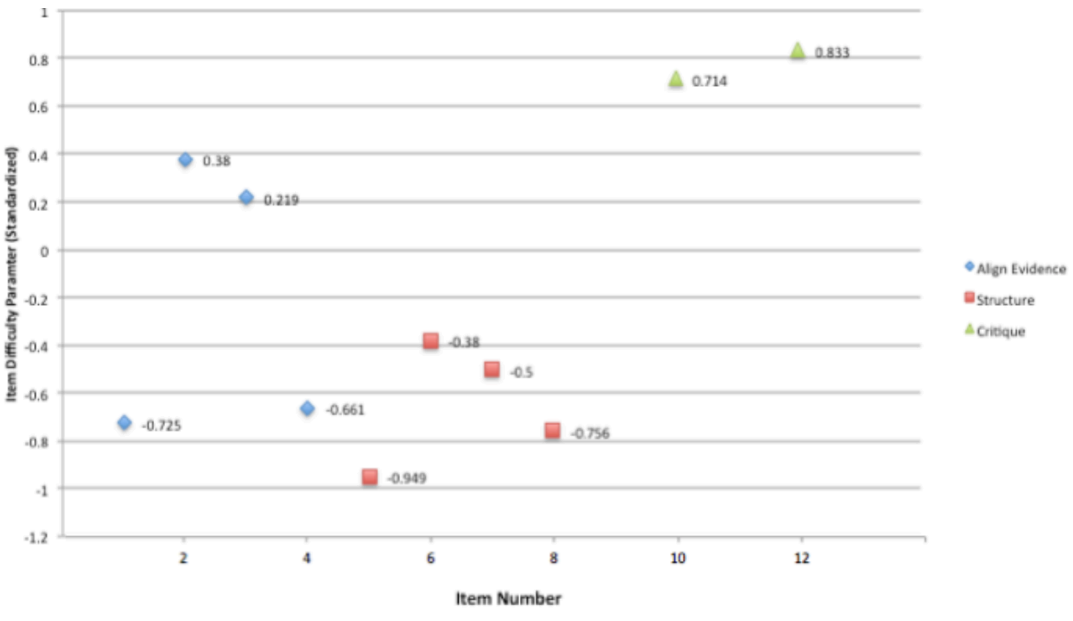
shown in figure 4.15, this includes the item discrimination parameters and item difficulty parameters for each item. Based on the 2PL model, the item difficulty (b) parameters were used to compare the relative difficulties of each item. The coefficients form the IRT model follow what was expected based on the Osborne et al. (2016) learning progression of argumentation. The critique items were the most difficult with item difficulty parameters of .714 and .833 for both critique items. This would mean that a student with an ability level (argumentation competency) of .833 standard deviations above average would have a 50/50 probability of answering question 12 correct. This means that when interpreting item difficultly parameters the higher values indicate a more difficult question, and lower values are easier questions.

Figure 4.15: 2PL model coefficients (a: discrimination parameter, b: difficulty parameter for argumentation assessment

Item #	Arg. Dimension	a	b
1	Align Evidence	1.174	-0.725
2	Align Evidence	1.222	0.380
3	Align Evidence	0.439	0.319
4	Align Evidence	0.874	-0.661
5	Structure	1.118	-0.949
6	Structure	1.135	-0.380
7	Structure	0.581	-0.500
8	Structure	0.677	-0.756
10	Critique	2.301	0.714
12	Critique	1.006	0.833

While there is a little overlap, the argument structure questions were the easiest with most of these items having the lowest item difficulty parameters. This leaves the evidence alignment questions as an intermediate in terms of difficulty. Based on the results on this assessment, these dimensions of argumentation do vary in difficulty, with critique being the most difficult, structure being the easiest, and alignment of evidence being a intimidate difficulty. The IRT parameters for the argumentation assessment are also displayed in Figure 4.16. Figure 4.16 plots the item difficulty parameter on the vertical axis, and the item number on the horizontal axis. The data label at each point indicates the items difficulty parameter, and the color of the point corresponds to the argumentation dimensions associated with the item.

Figure 4.16: Difficulty of items on the argumentation assessment (green=critique; blue=align evidence; red=structure)



Model Fit

Model fit is a measure of how well the 2PL IRT models accounts for the variance seen within the data set. Specifically, item fit is important because it can be an approximation for the confidence level of the estimation of IRT parameters. For this analysis, model fit was evaluated using the root mean square error approximation (RMSEA) with the criteria of the value being less than .05, and a comparative fit index (CFI) value of greater than .95. The results indicate that the 2PL has an acceptable level of model fit (RMSEA=.012; CFI=.981), suggesting that 2PL model adequately fits the argumentation assessment data.

RQ3: Relationship between content knowledge and argumentation dimensions

The purpose of this research question is to explore the extent to which students' level of content knowledge impact their ability to engage in specific argumentation dimensions. These argumentation dimensions include the alignment of claim and evidence, argumentation structure, and critique. Additionally, this research question allowed an exploration of the extent to which discipline specific knowledge impacts student performance on the various argumentation dimensions. It was intended to explore this topic using SEM methods, coupled with a factor analysis. However, as reported above the results of the factor analysis of the argumentation assessment is a one-factor model, and the items do not load to multiple factors that could be indicative of the argumentation dimensions. As such based on the data collected, this dissertation will not be able to explore research question. The following chapter of the dissertation will include a lengthy discussion in the limitations section about how to modify the current argumentation assessment to make it suitable for this type of analysis.

Chapter 5 Discussion

The purpose of this chapter is to provide a discussion of the results presented in the previous chapter, while also including a discussion of the implications of these results to the field of science education. As such, this chapter is divided into three main sections. The first section provides an in-depth discussion of the results and implications for each research question. Next, the second section provides a discussion of the limitations of this research study. The third section concludes the dissertation with the discussion of possible directions for future research.

Discussion of Results

This section has been organized around the research questions of this study. The subsections below will both summarize the results and discuss the findings in relation to current literature for each research question. This discussion will also focus on the implications of the results to both research in science education and the teaching and learning of science.

RQ1: The relationship between science content knowledge and science practice

The purpose of this research question was to explore the extent to which students' level of content knowledge can relate to their ability to engage in scientific argumentation. Researchers have previously examined the extent to which specific content knowledge can impact students' ability to construct and critique arguments (Koslowski, 1993; Lawson, 2003), and the predominant finding is that students require some level of content knowledge in order to be able to successfully create or critique an argument about that subject. However, this previous work has focused on content knowledge in a general sense, focusing on content knowledge from a broad domain level

perspective (Hogan & Maglienti, 2001; Zohar & Nemet, 2002). Given the previous work, this research question aimed to provide a more specific view of content knowledge by focusing on the impact of both domain general content knowledge and discipline specific content knowledge as defined by Alexander (1991) to determine if the level of content knowledge has an impact on students' ability to engage in scientific argumentation.

With the intention of exploring the impact of both domain general and discipline specific content knowledge on scientific argumentation, factor analysis, and structural equation modeling were used. Both the argumentation and water systems content assessments were subjected to factor analysis, and the resulting factor loadings were used to formulate the test model for SEM.

Results of Factor Analysis: Water Systems Content Assessment

Factor analysis was conducted on both the argumentation and water systems content assessments. The results of the factor analysis for the water systems content assessment suggested that the instrument was comprised of two factors. One of these factors can be conceptualized as evaporation knowledge or discipline specific knowledge with respect to the argumentation assessment as the items that relate to evaporation load onto this factor. The second factor on the water systems content test can be generally viewed as domain general water systems content knowledge. The factor identification was based on the variety of items that loaded onto the factor. The items that loaded onto this factor contained a variety of water systems topics including surface water, water cycle, groundwater, and watersheds. The topics of items loading into this factor are consistent with the water systems knowledge framework described by Gunckel et al. (2012), which includes a broad range of water systems ideas.

The results of the factor analysis are quite interesting in that there is a separate factor for evaporation knowledge, but not for other water systems topics. One possible explanation for this result was the imbalance of the number of evaporation items relative to the others. After removing poorly performing items, there were 18 items in the water systems test that were used in the factor analysis. Of these 18 items, there were far more evaporation items than there were items from each of the other water systems topics. As a result, one explanation for the factor analysis result is that the results are an artifact of the design of the water systems content test. In order to explore this possible explanation a further factor analysis was conducted on the water systems content test. However, for this analysis, some evaporation items were removed leaving only three of the seven evaporation items on the assessment. This was done to make the number of evaporation items on this test consistent with the number of items relating to the other water systems topics. The three evaporation items retained were randomly selected. After conducting the factor analysis, the results were not very different in that the factor analysis of the instrument with all seven evaporation items included, and the instrument was still comprised of two factors. One factor had the three evaporation items load onto it, while the other factor had the same water systems items load on it. Based on these results, the identification of an evaporation factor does not appear to be related to the differences in the number of items on the water systems test. These results were unexpected, as the water systems knowledge framework (Gunckel et al., 2012) does not include evaporation as being distinct from the rest of water systems topics. In addition to the number of items, there could be other explanations of this finding. It is possible that since all of the

evaporation items were given in succession at the end of the test, this could have impacted the ways in which student responded to the items.

Additionally, it is important to point out that four of the seven evaporation items on the content assessment were related to the rate of evaporation with two items being nearly identical. The remaining evaporation items involve examples of evaporation as well as an item pertaining to the definition of evaporation. Alternatively, the presence of an evaporation factor could be attributed to the nature of evaporation knowledge. Since evaporation was the only topic on the assessment that related to the subject of phase changes, perhaps this evaporation factor could be expanded if additional items were included that referenced other phase changes (i.e. condensation, sublimation), and this factor could be related to students' knowledge of phase change which may be discrete for knowledge of specific water systems. The water systems framework developed by Gunckel et al. (2012) subdivides the construct of water systems into its various environmental systems (i.e. surface systems, groundwater, atmospheric water, and the biotic system) and does not include phase changes as a specific component of the framework. However, it does include arrows connecting the different water systems that represent phase changes or the transfer of water between the systems. It may be that phase change knowledge taps into student knowledge about different water systems and in that regard is conceptually different for students than the other water systems. Given the results of the factor analysis, further research and work is needed to better understand how students conceptualize evaporation within water systems knowledge.

Results of Factor Analysis: Argumentation Assessment

The results of the factor analysis suggested that the argumentation assessment was comprised of a singular factor, and all of the items on the assessment loaded onto that one factor. This result is consistent with the research of several within the field (i.e. Osborne et al., 2016; Sampson & Clark, 2008; Ford, 2008). Despite the fact that the argumentation instrument contained three types of items (critique, alignment of evidence, structure), the results of the factor analysis suggested that the instrument is comprised of a singular factor for scientific argumentation. This singular argumentation factor was loaded on by items that covered several aspects of argumentation including the identification of argument structure, the alignment of claim and evidence, and critique. There are several explanations for the results of the factor analysis of the argumentation assessment, and they will be covered with more detail in the discussion of the results of the third research question.

Results of SEM. After running the factor analysis for both the water systems content and argumentation assessments, the results of these analyses were used to build the structural model that would be used in the SEM analysis. The results of the SEM analysis showed that the discipline specific knowledge or the evaporation latent variable had a stronger association with the argumentation latent trait than the domain general or general water systems latent trait. Since the SEM model is standardized, it suggests that an increase of one standard deviation of students' evaporation knowledge is associated with an increase in .85 standard deviations in argumentation knowledge. Additionally, the model suggests that an increase of one standard deviation in students' general water systems knowledge is associated with a .16 standard deviation increase in the students'

argumentation latent variable. While interpreting the SEM model, it is important to remember that looking at one linkage alone is not particularly useful as the domain general and discipline specific knowledge variables are latent and therefore they have no scale (Kline, 2016), and these linkages are interconnected in a larger and more complex model. However, it is more useful to compare the path coefficients of multiple linkages in the model that connect to the same latent variable (i.e. the association of domain general knowledge on argumentation v. the association of discipline specific knowledge on argumentation).

There are several explanations for the finding of the SEM model that discipline specific knowledge has a stronger association with argumentation than domain general knowledge. Based on this result it is logical to conclude that engaging in argumentation requires some level of content knowledge. This finding is consistent with the work of several other researchers in the field of scientific argumentation (Lawson, 2003; Sadler & Donnelly, 2007; von Aufschnaiter et al., 2008; Zohar & Nemet, 2002). However, previous work in this area had treated content knowledge as a broad construct and often base students' content knowledge off of performance on an assessment that focuses on domain general knowledge. There has not been work that specifically looked at the impact of discipline specific knowledge on scientific argumentation. It is here that this study is able to add to the current literature. Based on these results, students' knowledge about content that is closely related to the subject of the argument has a greater impact on their ability to engage in argumentation. While this issue has not been explored with regards to scientific argumentation, there has been some recent work that has discussed the influence of domain general knowledge on the use and analysis of evidence. Duncan,

Chinn, & Barzilai (2018) argued that the critical evaluation of evidence and claims require specific knowledge of the discipline, and often lay people struggle with this task due to an inadequate level of content knowledge. This discussion echoes the sentiment of the results of the SEM in that discipline level content knowledge is needed for engagement in argumentation.

Implications of Results

There are several implications of the results of the SEM analysis to both research in science education as well as the teaching and learning of science. When looking at the implications for research, there are specific implications in regard to the design of argumentation-based assessments. It is the goal of researchers to design valid and reliable assessments of scientific argumentation. Traditionally, assessments of scientific argumentation have been developed to be content light (Osborne et al., 2016). Essentially, this refers to the nature of the argumentation assessments being designed to include as little content knowledge as possible. While initially this sounds like a productive approach in the sense that it is ideal for an instrument measuring students' argumentation competency to only measure their ability to engage in argumentation. However, in practice, this is not an easy task to achieve. As the results of this study and the literature suggest, the ability for students to engage in argumentation requires some level of content knowledge. As such, assessment of argumentation competency inherently measures some degree of content knowledge for students to be consistently successful. Content knowledge is useful as students interpret the context of the argument (Berland & McNeil, 2010), and as the results of this study suggest students with a higher level of content knowledge are more successful at argumentation. While argumentation

assessments can never be completely removed of content knowledge, steps can be taken to reduce the extent to which students must rely on their own content knowledge to interpret the context of an argumentation assessment scenario. This reflects the NGSS orientation to viewing content and practice as inter-related. Given the interconnected nature of science content and science practice, assessment of the NGSS should be designed to be multi-dimensional. Overall, the results of this study highlight the difficulty in designing content light assessment given that the practice of argumentation is so closely related to the context of the argument.

In addition to the implications of the SEM results to research, there are also implications of these results to the teaching and learning of science in the classroom. Previous research has investigated the question of whether poor argumentation performance is a result of a lack of general competency or a lack of content knowledge (Hogan & Maglienti, 2001; Koslowski, 1996; Lawson, 2003), but there exists no definitive answer within the literature. The results of this study provide additional insight into the answer of this question. High performance on the water systems content assessment was associated with high scores on the argumentation assessment. This suggests as with other work in the field (i.e. Sadler & Donnelly, 2007; von Aufschnaiter et al., 2008; Zohar & Nemet, 2002) that students do indeed require some degree of content knowledge to engage in argumentation. Furthermore, the results of this study indicate that students need discipline specific knowledge about the phenomena they are arguing about more than tangentially related knowledge about the topic of the argument. This has implications to the classroom in that students struggling with argumentation may in fact actually possess the competencies needed to successfully argue, but they may lack

the content knowledge required. This sentiment is also echoed in the NGSS as within the standards content and practice are viewed as two interconnected constructs. As such, instruction around scientific argumentation can benefit from being embedded within content. If instruction of scientific argumentation occurs in the absence of content knowledge, then students may encounter difficulty arguing due to insufficiencies in their understandings of science phenomena or lacking the appropriate context to engage in argumentation. Thus, an approach to argumentation instruction with embedded content knowledge could be beneficial for students as they learn to engage in argumentation from evidence.

RQ2: Varying Difficulty of Argumentation Assessment Items

The purpose of this research question was to understand how different dimensions of argumentation vary in difficulty. Research into how students learn argumentation has resulted in different learning progressions of scientific argumentation (i.e. Berland & McNeil, 2010; Osborne et al., 2016). These learning progressions have established varying aspects of argumentation. Such aspects include critique, analysis of evidence, and argument structure. This research question explored the relative difficulties of these components of argumentation. Specifically, this research question explored the extent to which assessment items of different argumentation dimensions varied in difficulty. This question explored the following dimensions: argumentation structure, alignment of evidence, and critique. Based on the learning progressions of argumentation (i.e. Osborne et al., 2016) critique was expected to be the most difficult dimension, with alignment of evidence being easier, and argument structure being the easiest dimension.

In order to determine the relative difficulties of the assessment items, the argumentation assessment was fitted with a 2PL IRT model, and the subsequent IRT parameters were estimated. Specifically, the results of this research question pertained to the estimated item difficulty parameter for each item. The difficulty parameter provides a standardized measure of item difficulty which was used determine the relative difficulties of the items relating to different dimensions of argumentation.

Result of IRT analysis

The results of the IRT analysis showed some separation of the argumentation dimensions based on their difficulty. Of the three dimensions of argumentation, critique was the most difficult, with alignment of evidence and argumentation structure being slightly clustered. Overall, argumentation structure questions were easier than alignment of evidence. These results are consistent with previous literature on scientific argumentation. Critique was the most difficult dimension of augmentation, which is consistent with previous learning progressions of argumentation (Berland & McNeil, 2010; Osborne et al., 2016). Researchers have also made a case for the increased difficulty of critique by leveraging the cognitive demands of critique in comparison to other argumentation dimensions. Ford (2008) argues that the ability to critique requires different competencies than building an argument. Critique requires the ability to identity what the claim is in an argument, determining what constitutes the data used to support the argument, and also what kind of reasoning is used to explain how the evidence supports the claim. In the argumentation assessment used in this study, the critique items require students to consider two competing arguments, and determine why one argument was better than another one. This type of task requires cognitive skills in comparing and

contrasting the relative merits of two arguments simultaneously, which is cognitively more demanding than identifying the parts of an argument or determining which argument a piece of evidence supports. For these reasons, the critique items were expected to be the most difficult items on the argumentation assessment, which they were.

In addition to critique items, the argumentation assessment also included alignment of evidence and identification of argument structure items. Items pertaining to the alignment of evidence presented students with two competing claims, and asked students to determine which of two competing claims a piece of evidence supported. While there was some overlap with argument structure, alignment of evidence was generally more difficult than structure, but easier than critique. This finding is also consistent with the literature on scientific argumentation. Researchers (i.e. Driver, Newton, & Osborne, 2000; Sampson & Clark, 2008) have noted that argumentation is a difficult competency for students to learn. Specifically, the ability to students to determine which competing idea that a piece of evidence supports requires students to draw on content knowledge in addition to their knowledge of arguments (Osborne et al., 2016). As a result, according to the Osborne et al. (2016) learning progression of augmentation, alignment of evidence requires more degrees of coordination and thus should be more difficult than simply identifying the various parts of an argument (i.e. claim, evidence and reasoning).

Again, while there was some overlap of the results of the IRT analysis between the dimensions of alignment of evidence and structure, it likely that these dimensions would have been more clearly separated if the dimensions of evidence alignment did not

have such a large degree of variation in difficulty. The other two dimensions (critique, structure) both have item difficulties that are relatively clustered together. However, alignment of evidence exhibits a sizeable amount of variation in the estimated difficulty parameters. When looking closer at the argumentation assessment item difficulty plot, two of the four alignment of evidence items are relatively easy, while the other two are much more difficult. The two easier items are cases where the correct answer is that a piece of evidence only supports one of the competing arguments, but cases where pieces of evidence supported both of the competing arguments were much more difficult for students. Given this result, it would interesting to follow-up on this finding and include more alignment of evidence questions on a future version of the argumentation assessment and see if this pattern hold true for a larger number of items, as the present study only includes four items in the alignment of evidence dimension.

In summary, the results of the IRT analysis for the argumentation assessment were that the identification of argument structure was the easiest items on the assessment. Alignment of evidence items were more challenging and in particular aligning evidence that supports more than one competing claim was harder for students than aligning a piece of evidence to a single claim. The most difficult items on the assessment were the critique items. As noted above, these findings are consistent with several documented learning progression for scientific argumentation (i.e. Berland & McNeil, 2010; Osborne et al., 2016).

Implications of Results

The results of the IRT item difficulty analysis of the argumentation assessment has clear implications for both research in the area of argumentation, and the teaching

and learning of argumentation in classrooms. When looking at the research implications, these relate to the way in which argumentation can be assessed. A current view of scientific argumentation within the literature is that argumentation is a singular construct that is made up of several competencies (Osborne et al., 2016; Sampson & Clark, 2008). In that regard, the results are consistent with previous literature in that the instrument used in this study was unidimensional, but contained various competencies that varied in difficulty. As noted previously in this dissertation, there are a wide variety of ways that argumentation has been assessed within science education. These assessments range from written arguments (Osborne et al., 2016; Sampson, Enderle, Grooms, & Whittle, 2014; Kelley & Takao, 2002) to classroom discourse (Chin & Osborne, 2010; Hogan, Nastasi, Pressley, 1999) of oral arguments being made between students. While these methods of assessing student's ability to argue allow students to openly form their own arguments, they are also much more labor intensive to score and evaluate. Comparably, there are much fewer assessment of argumentation that are multiple-choice in nature, and allow students to engage in various argumentation competencies. To this end, the argumentation assessment used in this study is multiple-choice, but also allows students to engage in critique, align evidence with a claim, and identify argumentation structure. Additionally, this instrument also provides a fairly reliable measure of argumentation competency as noted by both the classical test reliability (α=.68), as well as the IRT information function. As such, this instrument provides a measure of argumentation that is reliable and easy to score due to the multiple-choice format of the instrument.

In addition to research implications, the results of the IRT analysis also have implications to the teaching and learning of argumentation. As noted in the discussion of

the first research question, instruction of argumentation should not be isolated from content because in order to for students to engage in argumentation they require some level of content knowledge about the topic of phenomenon they are arguing about. With that said, the results of the IRT provide an understanding of the relative difficulties of various argumentation competencies. As such, argumentation instruction should begin with easier competencies such as identifying argument structure, and build to include more complex and cognitively demanding competencies such as critique. The notion of difficulty of argumentation dimensions is not novel to this study, and similar findings have been reported and included in various learning progression of scientific argumentation (Berland & McNeil, 2010; Osborne et al., 2016).

RQ3: The relationship between science content knowledge and the dimensions of argumentation

The purpose of this research question was to explore the connection between students' level of content knowledge and the various dimensions of competencies of the argumentation construct. This allows for a deeper understanding of how students apply content knowledge into different competencies of argumentation rather than looking at the entire argumentation construct as a whole. For instance, content knowledge may be more important for certain argumentation tasks (i.e. critique) and less important for others. In this regard, this question builds on the finding from the first research question that students' discipline specific knowledge has a stronger association with argumentation performance than does domain general content knowledge. The investigation of this research question allows for the exploration of discipline specific and

domain general content knowledge to be extended to analyze their associations with the various competencies of the argumentation construct.

Results of data analysis

The data analysis procedure for this research question was similar to the analysis approach used to explore the first research question. This procedure involved conducting an exploratory factor analysis on both the water systems content and argumentation assessments, and based on the results of those analyses use structural equation modeling to create a structural model that estimates the relationships between content knowledge and argumentation. However, this study was not able to explore this issue due to the results of the factor analysis of the argumentation assessment. The results of the factor analysis suggested that the argumentation assessment was a unidimensional instrument, and thus there was not empirical justification for breaking down the argumentation assessment into more than one factor. Ideally, the argumentation assessment would have contained three factors, each of which was loaded onto by items that related to the three different argumentation competencies (critique, alignment of evidence, argument structure). There are multiple explanations for the result of the factor analysis. First, it could be the construct of argumentation is truly unidimensional, and all of the argumentation competencies actually map onto a singular construct that is scientific argumentation. This stance is supported with the literature as researchers have viewed scientific argumentation as being comprised of a singular dimension (Lee et al. 2014; Sampson & Clark, 2008). With that being said, the finding from the second research question of this study is interesting. Based on the IRT item analysis there seemed to be at some level a clustering pattern of item difficulties. While the clustering was not

conclusive and there was certainly overlap between some of the alignment of evidence items and the argument structure item, critique items were clearly the most difficult. As such, another explanation of the factor analysis results could be that the construct of argumentation is actually multidimensional, but this was not reflected in the factor analysis results due to measurement constraints of the argumentation assessment. The argumentation assessment instrument was made up of ten multiple-choice items that covered three broad argumentation competencies. These competencies included identification of argument structure, alignment of evidence, and critique. Since there were only ten items on this assessment, there were only four argument structure items, four alignment of evidence items, and two critique items. As such there is the possibility that result of the factor analysis showing the presence of a singular argumentation factor may be attributed to the low number of items for each argument competency. Kline (2016) made the suggestion that generally when conducting a confirmatory or exploratory factor analysis it is advised to have a minimum of three variables per factor with generally strong loadings (above .4). While this minimum criterion was achieved for two of the three factors, the loadings were not as strong as desired. The minimum variable number recommendations of factor analysis were met for both the alignment of evidence and argumentation structure, but the in both of those cases not all of the item loadings were above .4. Furthermore, both alignment of evidence and argument structure had two item loadings above the .4 recommendation, and two items with loadings below this value.

Based on these results there are a couple of recommendations that could be made to this instrument in the future to better determine if the removal of measurement constraints of the argumentation assessment would allow for multiple factors to be

retained from a factor analysis. The main recommended changes to the argumentation assessment include the addition of more items that relate to the argumentation competences that are measured on the assessment with a goal of six to seven items per argumentation competency. This is an estimated range and it is important not to have too many items, as students are already required to read and analyze several arguments through the course of this assessment. However, this change is especially needed for critique, as there are currently only two critique items on the assessment. While the assessment originally had four critique items, two were dropped since they asked students to select a better argument from two possible choices. Since these questions only had two answer choices, they were extremely easy and, based on the IRT analysis, the items were not able to discriminate at all based on student ability. As such, it is imperative that any additional critique items require students to analyze why one argument is better than another much like the two critique items that were included in this analysis. Additionally, adjustments to the argument structure items might include more instances of students being asked to identify pieces of an argument. Ideally, there would be an equal number of items asking students to identify an arguments' claim, evidence, and reasoning. The present version of this assessment has four argument structure questions. In the current version students are presented with two competing arguments, and are asked to identify the claim for both, and the evidence and reasoning of just one. As such, it is reasonable to add two structure items that require students to identify the reasoning and evidence for both arguments. Making such a change would also not increase the number of arguments that students have to analyze, which is desirable. Finally, changes should be made to the alignment of evidence portion of the assessment. Currently, this section of the assessment

provides students with two competing claims and four pieces of evidence. Students are asked to determine which argument each piece of evidence supports. Adding two pieces of evidence will allow for more items on the assessment that require students to align evidence with claims. Additionally, it may be beneficial to add pieces of evidence that support only one of the competing claims, as the IRT analysis suggested that the it is more difficult for students to align pieces of evidence that support both or neither of the presented claims.

In summary, the relationship between content knowledge and argumentation competencies (i.e. alignment of evidence, structure, critique) could not be explored in this dissertation due to the factor analysis of the argumentation assessment being composed of a singular factor. This could be due to inherent property of the argumentation construct in that it is unidimensional, or it could be caused by measurement constraints of the argumentation instrument used in this analysis. A way to address this issue is to modify the argumentation assessment to include more items that relate to each argumentation competency as the minimum number of variables suggested for each factor (i.e. Kline, 2016; Pituch & Stevens, 2016) was not met for each of the three argumentation competencies.

Limitations

The purpose this section is to acknowledge and explain the limitations of this dissertation. First, it is important to discuss the limitations of the assessment instruments. When looking at both of argumentation and water systems content assessment, it is important to point out the IRT information functions are not perfect. Specifically, when analyzing the information function of the water systems assessment, it is clear that the

instrument does not provide a desired amount of information for students of higher ability. The information curve suggests that the water systems content assessment is reliable for students of below average to about one standard deviation above average water systems knowledge. This is likely due to the water systems assessment containing several easier items, and as a result the assessment provides more information for students that have lower than average water systems knowledge.

Additionally, the argumentation assessment also has some limitations in relation to its reliability. When looking at the information function for the argumentation assessment, it is clear that there are areas where reliability could be improved. Here we see that the argumentation instrument is most reliable for students that have slightly higher than average argumentation ability. This is likely due to the high number of difficult argumentation items on the assessment, but also in part attributed to the innate difficulty of engaging in scientific argumentation, as many scholars in the field have noted the high difficulty of argumentation (Osborne et al., 2016; Sampson & Clark, 2008). With that being said, both the argumentation (.68) and water systems content (.72) assessments have an acceptable level of marginal reliability. Marginal reliability is based on the true score model (Lord & Novick, 1968) and is an estimate of the overall reliability often closely in value to the alpha coefficient (De Ayala, 2016).

In addition to the reliability of the assessments, there were additional limitations to the argumentation assessment that are worth discussion. The design of this study uses a multiple-choice assessment of scientific argumentation and thus does not allow students to construct an argument. As such, the assessment used in this study is more of a measure of students being able to engage critically in various dimensions of argumentation (i.e.

critique, argument structure, and alignment of evidence), but the argumentation assessment does not require students to generate an argument. Also, by nature of the multiple-choice format, it is unclear if students engaged in the same cognitive processes that would be expected of argumentation. One approach to answering this question would be to conduct an analysis of cognitive validity for the argumentation assessment. Kane (1992) outlines an argument based approach to instrument validity. In this discussion, Kane described methods to obtain a better understanding of the types of cognitive processes that are involved with responding to test items. Further elaboration of such methods includes conducting interviews with test respondents to determine if the items required the cognitive skills that were expected (Kane 1992; Kane, 2001). Essentially, this refers to ensuring that the assessment items engage the cognitive processes of the respondents that were expected and intended by the assessment designers. Specifically, this is relevant to the process of critique. Based on the argumentation literature, most researchers agree that engaging in critique is a cognitively demanding task. Based on the this recommendation from Kane, conducting brief interviews with students who completed the argumentation assessment could improve understanding of the types of cognitive processes elicited via the assessment instrumentation, This analysis could then be used to determine if the cognitive demands of specific items were consistent with what would be expected based on the scientific argumentation literature. This information could help inform the validity of the instruments beyond the construct validity that was established in this dissertation.

Further limitations of the assessment include the measurement scale. As noted within the methodology chapter, the assessment data used in this analysis was

dichotomous in nature. Student data for specific items were scored as either 1 (correct) or 0 (incorrect), and this was then used for the exploratory factor analysis and the subsequent SEM analysis. While the principal axis factoring method was used in this study for the exploratory factor analysis since it lessens the statistical assumption of multivariate normality, there is no denying that the use of dichotomous data removed some of the nuance of student argumentation competency and could have potentially contributed to the result of the argumentation instrument retaining only a single factor. Alternatively, student logit data could have been used from the IRT model, but it is likely that this would not have drastically changed the SEM outcome. Regardless of the results of the factor analysis of the argumentation instrument, several SEM models were created that included multiple argumentation latent variables and all of these model failed to converge. This indicated that a model with multiple argumentation latent variables was an overall poor fit with the water systems content data.

In addition to the assessment instruments, there were also limitations to the SEM analysis. While the intention of this dissertation was to explore the relationship between argumentation and content knowledge, it is important to acknowledge that the structural model fails to account for epistemic knowledge. Sandoval & Reiser (2005) discussed epistemic knowledge as knowledge about the kinds of questions that can be answered through investigation, the methods that are accepted for generating data, and an understanding of what counts as legitimate interpretation of data. Particularly the aspects of epistemic knowledge that relate the determination of what counts as evidence and identification of accepted methods for generating data are directly applicable to several aspects of argumentation. For example, when constructing an argument students leverage

epistemic knowledge in selecting appropriate evidence to support a claim. Additionally, epistemic knowledge is used when students engage in critique (Ford, 2008). This limitation is due to the absence of an assessment marker for epistemic knowledge, but based on the literature epistemic knowledge is a significant factor in argumentation. Sandoval's framework for examining arguments placed emphasis on evaluating the epistemic quality of arguments (Sandoval 2003; Sandoval & Millwood, 2005). Furthermore, Sandoval (2003) and Duschl (2008) have argued that the pursuit of epistemic goals and the establishment of epistemic criteria for the evaluation of science claims should become a core component of argumentation practices. As such, this highlights the importance of epistemic knowledge on argumentation, and the SEM model stands to improve its estimation of students' argumentation ability by adding a measure of epistemic knowledge.

Finally, it is important to reiterate that this study only examined the relationship between argumentation and content knowledge within one specific water systems topic. The argumentation assessment only allowed students to engage in argumentation around the topic of evaporation, and as such the findings of the study are all situated within argumentation related to evaporation. While it is likely that the observed results would be consistent across several content areas, more data would need to be collected in order to verify this notion. Specifically, this would require the development of a new argumentation instrumentation that was situated within various content domains, as well as the use of different instrumentation to assess students' level of content knowledge.

Directions for Future Research

This chapter concludes with a discussion of future areas of research. While the third research question of this study aimed to explore the relationship between content knowledge on the various competencies of argumentation, this question remained unanswered due to the unidimensional nature of the argumentation assessment. As suggested previously in this chapter, the unidimensional nature of the argumentation assessment could be due to constraints in the argumentation assessment and several modifications could be made to the argumentation assessment to remove measurement constraints. Future research could conduct similar SEM analysis using a revised argumentation assessment with more items relating to each of the argumentation competencies to potentially allow for a multi-dimensional instrument that can parse out critique, alignment of evidence, argument structure and analyze these using SEM in context of the water systems content assessment. Perhaps with these modifications, the argumentation assessment could contain multiple factors that could be used to further explore the relationship of content knowledge and argumentation by better understanding which dimensions of argumentation are more reliant on content knowledge

Additionally, this study explored the connection between science content and science practice in regards to water systems. Future work could expand to include several other science phenomena. Specifically, it would be interesting to explore science concepts that are more abstract for students such as the dissolving of materials or some phenomena that students cannot directly observe. Such work could help determine if students rely on content knowledge when arguing about concepts that are not directly observable (i.e. molecular biology, chemistry). In the case of this study, the

argumentation scenario was in the context of evaporation, and the evidence used was primarily based on a direct observation (i.e. after some time water in bowl disappears) that students could easily see and identify. Perhaps when arguing about more abstract phenomena there may be an even stronger association between content knowledge and argumentation than what was reported in this dissertation.

Along the same lines, future work building off of this dissertation could explore the relationship between content and practice on a broader scale. The first chapter of this dissertation makes an argument for the practice of argumentation being a representative science practice since competency in argumentation requires competency in several other practices (i.e. analyzing and interpreting data, explanation of science phenomena). Despite argumentation being a representative practice, engaging in different science practices could tap into content knowledge in different ways. As such, the types of analyses used in this dissertation would be suitable for the exploration of the relationship between content knowledge and other science practices such as modeling, analysis of data, or generating explanations for science phenomena.

The completion of this dissertation has provided insights into several aspects of science education. Broadly speaking, the results of this study provide insights into the nature of scientific argumentation, as well as the design of NGSS aligned assessments. As noted earlier in this dissertation, the NGSS represents a significant departure from previous science standards. The NGSS aimed to support and encourage students' ability to explain science phenomena and solve problems through an integration of science and engineering practices, disciplinary core ideas, and crosscutting concepts. The unique approach of the NGSS promotes development of students' ability to explain scientific

phenomena and design solutions to problems through their engagement with science practices. Given the importance that the NGSS places on science practice, it is important to better understand the skills and knowledge students need in order to successfully engage in science practices. The results of this study indicate that students with a deeper understanding of science content are more successful at engaging in science practices. This finding can help inform the design of instruction of science practices by providing more detailed information about the knowledge that students need in order to be successful and ultimately reach the ambitious goals set forth by the NGSS.

As noted earlier, the NGSS promotes a deeper understanding of both science content knowledge and the process of science itself. This shift is extremely ambitious, and requires valid and reliable assessment instrumentation to evaluate the extent to which students reach these ambitious goal set forth by the NGSS. To this end, the work of this dissertation provides additional insights into the ways in which the dimensions of NGSS can be assessed. Namely, the results of this study suggest that the close relationship between science content and science practice highlights challenges to the design and development of assessments that aim to measure student competency in the various science practices. These challenges stem from the close relationship between science content and science practice, and thus assessments of science practices will inevitably also assess students' level of content knowledge at some level. While this was partly understood from previous research, the results of this study indicate that students' level of discipline specific content knowledge has the greater impact on science practices. As a result, the work of this dissertation has helped to better understand the knowledge that students leverage when engaging science practices. Better understanding the knowledge

that is required for students to engage in science practices can help assessment developers create instrumentation that are more effective at assessing science practices and the learning of NGSS in general.

In conclusion, this dissertation provided for a better understanding of the relationship between content knowledge and science practice by viewing content knowledge more specifically to include domain general and discipline specific content knowledge. The results of this study provide researchers with a better understanding of the nature by which students engage in argumentation, and the ways in which students use content knowledge when engaging in argumentation. Further understanding the connection between content knowledge and argumentation has the potential to inform and improve argumentation instruction, which ultimately can provide students with more authentic science experiences in the classroom.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. Applied psychological measurement, 21(1), 1-23.

Alexander, P. A., Pate, P. E., Kulikowich, J. M., Farrell, D. M., & Wright, N. L. (1989). Domain-specific and strategic knowledge: Effects of training on students of differing ages or competence levels. Learning and Individual Differences, 1(3), 283-325.

Alexander, P. A., Schallert, D. L., & Hare, V. C. (1991). Coming to terms: How researchers in learning and literacy talk about knowledge. Review of educational research, 61(3), 315-343.

Alonzo, A. C., & Steedle, J. T. (2008). Developing and assessing a force and motion learning progression. Science Education, 93, 389–421.

Anderson, R.D. (2002). Reforming science teaching: What research says about inquiry. Journal of Science Teacher Education, 13 (1), p. 1-12.

Aydeniz, M., Pabuccu, A., Cetin, P. S., & Kaya, E. (2012). Argumentation and Student's Conceptual Understanding of Properties and Behaviors of Gases. International Journal of Science and Mathematics Education, 1-22.

Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. Written communication, 2(1), 3-23.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological bulletin, 107(2), 238.

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.

Bransford, J. D., Brown, A., & Cocking, R. (1999). How people learn: Mind, brain, experience, and school. Washington, DC: National Research Council.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. Educational Assessment, 11(1), 33-63.

Bybee, R., & McCrae, B. (2011). Scientific literacy and student attitudes: Perspectives from PISA 2006 science. International Journal of Science Education, 33(1), 7-26.

Callender, J. C., & Osburn, H. G. (1977). A method for maximizing split-half reliability coefficients. Educational and Psychological Measurement, 37(4), 819-825.

Cattell, R. B. (1988). The meaning and strategic use of factor analysis. In Handbook of multivariate experimental psychology (pp. 131-203). Springer, Boston, MA.

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

Crombie, A. C. (1996). Science, Art and Nature in Medieval and Modern Thought. London ; Rio Grande, Ohio: Bloomsbury Academic.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. psychometrika, 16(3), 297-334.

Cross, D., Taasoobshirazi, G., Hendricks, S. & Hickey, D. T. (2008). Argumentation: A strategy for improving achievement and revealing scientific identities. International Journal of Science Education, 30(6), 837–861.

De Ayala, R. J. (2013). The theory and practice of item response theory. Guilford Publications.

De Vries, E., Lund, K., & Baker, M. (2002). Computer-mediated epistemic dialogue: Explanation and argumentation as vehicles for understanding scientific notions. The journal of the learning sciences, 11(1), 63-103.

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

Duncan, R. G., & Gotwals, A. W. (2015). A tale of two progressions: On the benefits of careful comparisons. Science Education, 99(3), 410-416.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105(3), 399-412.

Duschl, R. A., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. Studies in Science Education, 47(2), 123-182.

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

Erduran, S., & Jiménez-Aleixandre, M. P. (2008). Argumentation in Science Education: Perspectives from Classroom-Based Research. Florida State University-USA: Springer.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. Science education, 88(6), 915-933.

Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. Structural equation modeling: A second course, 10(6), 269-314.

Flavell, J. H. (1987). Speculations about the nature and development of metacognition. Metacognition, motivation, and understanding, 21-29.

Ford, M. J. (2015). Learning progressions and progress: An introduction to our focus on learning progressions. Science Education, 99(3), 407-409.

Ford, M. J. (2008). Disciplinary authority and accountability in scientific practice and learning. Science Education, 92(3), 404–423.

Ford, M. J., & Forman, E. A. (2006). Chapter 1: Redefining disciplinary learning in classroom contexts. Review of research in education, 30(1), 1-32.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. Science Education, 94(2), 259–281.

Graves, M. F., Slater, W. H., & White, T. G. (1989). Teaching content area vocabulary. Content Area Reading and Learning, 214-224.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21(4), 347-360.

Greeno, J. G. (1989). A perspective on thinking. American Psychologist, 44, 134-141.

Gunckel, K. L., Covitt, B. A., Salinas, I., & Anderson, C. W. (2012). A learning progression for water in socio-ecological systems. Journal of Research in Science Teaching, 49(7), 843-868.

Hammer, D., & Sikorski, T. F. (2015). Implications of complexity for research on learning progressions. Science Education, 99(3), 424-431.

Hewson, P. W. (1985). Epistemological commitments in the learning of science:

Examples from dynamics. The European Journal of Science Education, 7(2), 163-172.

Hogan, K., & Maglienti, M. (2001). Comparing the Epistemological Underpinnings of Students' and Scientists' Reasoning about Conclusions. Journal of Research in Science Teaching, 38(6), 664–687.

Hogan, K., Nastasi, B. K., & Pressley, M. (1999). Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. Cognition and instruction, 17(4), 379-432

Holmes, B. C. (1983). The effect of prior knowledge on the question answering of good and poor readers. Journal of Reading Behavior, 15(4), 1-18.

Houts, C. R., & Cai, L. (2013). flexMIRT user's manual version 2.0: flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural equation modeling: a multidisciplinary journal, 6(1), 1-55.

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), 757-792.

Jin, H. & Anderson, C. W. (2012). A learning progression for energy in socio- ecological systems. Journal of Research in Science Teaching, 49(9), 1149- 1180.

Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university

oceanography students' use of evidence in writing. Science education, 86(3), 314-342.

Khine, M. S. (2012). Perspectives on Scientific Argumentation. Springer.

Kline, R. B. (1998). Software review: Software programs for structural equation modeling: Amos, EQS, and LISREL. Journal of psychoeducational assessment, 16(4), 343-364.

Kline, R. B. (2015). Principles and practice of structural equation modeling. Guilford publications.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. Journal of Psychology, 216(2), 61-73.

Koslowski, B. (1996). Theory and Evidence: The Development of Scientific Reasoning. MIT Press.

Krajcik, J., Codere, S., Dahsah, C., Bayer, R., & Mun, K. (2014). Planning instruction to meet the intent of the next generation science standards. Journal of Science Teacher Education, 25, p. 157-175.

Kuhn, D. (1991). The Skills of Argument. Cambridge University Press.

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

Latour, B., & Woolgar, S. (1986). Laboratory Life: The Construction of Scientific Facts (2nd ed.). Princeton, NJ: Princeton University Press.

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

Lee, M. H., Johanson, R. E., & Tsai, C. C. (2008). Exploring Taiwanese high school

students' conceptions of and approaches to learning science through a structural equation modeling analysis. Science Education, 92(2), 191-220.

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. Journal of Research in Science Teaching, 51(5), 581–605.

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

Lehrer, R., & Schauble, L. (2012). Seeding evolutionary thinking by engaging children in modeling its foundations. Science Education, 46(6), 731-735.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. ETS Research Report Series, 1984(1).

Lord, F. M., & Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mason, L., & Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. Learning and Instruction, 16(5), 492–509.

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

Metcalfe, J., & Shimamura, A. P. (Eds.). (1994). Metacognition: Knowing about knowing. MIT press.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning

progression for carbon cycling in socio-ecological systems. Journal of Research in Science Teaching, 46(6), 675-698.

Muthén, B., Du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. 1997. Technical Report.

National Research Council (1996). National Science Education Standards. Washington, DC: The National Academy Press.

National Research Council (NRC). (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: The National Academy Press.

NGSS Lead States. (2013). Next Generation Science Standards: for states, by states. Washington, DC: The National Academies Press.

O'Rourke, N., Psych, R., & Hatcher, L. (2013). A step-by-step approach to using SAS for factor analysis and structural equation modeling. SAS Institute.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. Journal of research in science teaching, 53(6), 821-846.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). Developing Assessments for the Next Generation Science Standards. National Academies Press. 500 Fifth Street NW, Washington, DC 20001.

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

Peterson, P. L. (1988). Teachers' and students' cognitional knowledge for classroom

teaching and learning. Educational researcher, 17(5), 5-14.

Pituch, K. A., Stevens, J., & Stevens, J. (2016). Applied multivariate statistics for the social sciences.

Plummer, J. D., & Krajcik, J. (2010). Building a learning progression for celestial motion: Elementary levels from an Earth-based perspective. Journal of Research in Science Teaching, 47(7), 768-787.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychological bulletin, 114(3), 552.

Reiser, B.J. (2013). What professional development strategies are needed for successful implementation of the next generation science standards? In the Invitational Research Symposium on Science Assessment presented conducted at The Center for K1- 12 Assessment and Performance Management at Educational Testing Services, Washington, DC.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal, 3(4), 369-379.

Sadler, T. D., & Donnelly, L. A. (2006). Socioscientific argumentation: The effects of content knowledge and morality. International Journal of Science Education, 28(12), 1463-1488.

Sadler, T., & Fowler, S. (2006). A threshold model of content knowledge transfer for socioscientific argumentation. Science Education, 90(6), 986–1004.

Sadler, T. D., & Zeidler, D. L. (2005). The significance of content knowledge for

informal reasoning regarding socioscientific issues: Applying genetics knowledge to genetic engineering issues. Science Education, 89(1), 71–93.

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.

Sampson, V., Enderle, P., Gleim, L., Grooms, J., Hester, M., Southerland, S., & Wilson, K. (2014). Argument-Driven Inquiry in Biology: Lab Investigations for Grades 9- 12. NSTA Press, Arlington, VA.

Shavelson, R. J. (2009). Reflections on learning progressions. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Shea, N. A., & Duncan, R. G. (2013). From theory to data: The process of refining a learning progression. Journal of the Learning Sciences, 22(1), 7-32.

Sikorski, T. F., & Hammer, D. (2010). A critique of how learning progressions research conceptualizes sophistication and progress. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), Learning in the disciplines: Proceedings of the 9th International Conference of the Learning Sciences (Vol. 1, pp. 1032–1039). Chicago, IL: International Society of the Learning Sciences.

Smith, C., & Wiser, M. (2015). On the importance of epistemology-disciplinary core concept interactions in LPs. Science Education, 99(3), 417-423.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. Journal of Research in Science Teaching, 46(6), 610–631.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. Journal of Research in Science Teaching, 46(6), 699–715.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. Multivariate behavioral research, 25(2), 173-180.

Strimaitis, A. M., Schellinger, J., Jones, A., Grooms, J., & Sampson, V. (2014). Development of an instrument to assess student knowledge necessary to critically evaluate scientific claims in the popular media. Journal of College Science Teaching, 43(5), 55-68.

Toulmin, S. (1958). The Uses of Argument. Cambridge: Cambridge University Press.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. Educational measurement: Issues and practice, 10(1), 37-45.

Venville, G. J. & Dawson, V. M. (2010). The impact of a classroom intervention on grade 10 students' argumentation skills, informal reasoning, and conceptual understanding of science. Journal of Research in Science Teaching, 47, 952–977.

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.

Voss, J. F., Blais, J., Means, M. L., Greene, T. R., & Ahwesh, E. (1986). Informal reasoning and subject matter knowledge in the solving of economics problems by naive and novice individuals. Cognition and instruction, 3(3), 269-302.

Voss, J.F., & Means, M.L. (1991). Learning to reason via instruction in argumentation. Learning and Instruction, 1, 337-350.

Walker, J., Sampson, V., Grooms, J., Anderson, B., & Zimmerman, C. (2010). Argument-driven inquiry: An instructional model for use in undergraduate chemistry labs. In Annual International Conference of the National Association of Research in Science Teaching (NARST).

Walker, C. H. (1987). Relative importance of domain knowledge and overall aptitude on acquisition of domain-related information. Cognition and Instruction, 4, 25-42.

Walton, D. N. (1990). What is reasoning? What is an argument? The Journal of Philosophy, 87(8), 399–419.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. Journal of Research in Science Teaching, 46(6), 716-730.

Wilson, M., & Draney, K. (2004). Some Links Between Large-Scale and Classroom Assessments: The Case of the BEAR Assessment System. Yearbook of the National Society for the Study of Education, 103(2), 132-154.

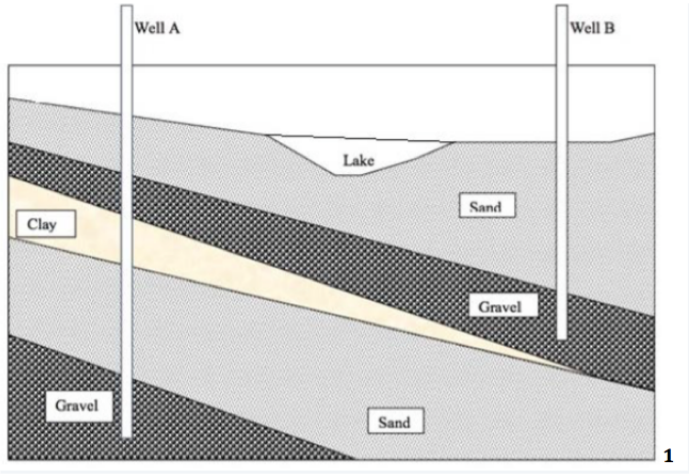
Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. Applied measurement in education, 13(2), 181-208.

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.

Appendix : Assessment Instrumentation

Water Systems Content Assessment

For the following items, refer to the image below, which represents a cross-sectional view of the underground space close to two wells.



1. Of the two wells in the picture above, which will likely have the cleanest water?

- ○ Well A
- ○ Well B
- ○ The same

2. Will water flow most easily through sand, gravel, or clay?

- ○ Sand
- ○ Gravel
- ○ Clay

3. Water will flow most slowly through which material?

- ○ Sand
- ○ Gravel
- ○ Clay

4. Which of the following is a major source of moisture that reaches or becomes part of the Earth's atmosphere?

- ○ Lakes
- ○ Rivers
- ○ Oceans
- ○ Polar ice caps

5. When the temperature of water and the atmosphere becomes colder, the rate of evaporation:

- ○ Decreases
- ○ Increases
- ○ Stays the same

6. From where does most of the Earth's water evaporate?

- ○ Lakes
- ○ The ocean
- ○ Trees

7. Most watersheds eventually drain into:

- ○ A river
- ○ A lake
- ○ An ocean

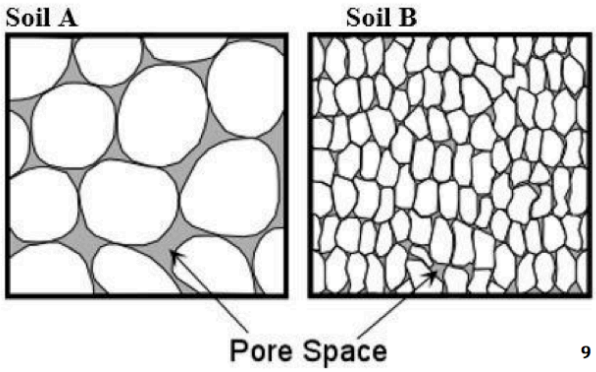
8. Where does the energy that powers the water cycle come from?

- ○ The ocean
- ○ Thunderstorms
- ○ The sun

9.Refer to the map of Missouri above. The Missouri River flows from Kansas City to Saint Louis. Predict which city has the higher elevation above sea level.

- ○ Kansas City
- ○ Saint Louis
- ○ They have the same elevation

For the next two items, consider the following diagram of two different soils.



10. Which soil will hold the most water?

- Soil A
- Soil B
- No difference

11. In science class students added water to both Soil A and Soil B. They found that water was able to move quickly through Soil A, but not Soil B. Why can water move quicker through Soil A than Soil B?

- Soil B is harder than Soil A
- Soil B is made up of smaller particles than Soil A
- Soil B has less space between the particles than Soil A

For the following items, refer to the topographic map below.

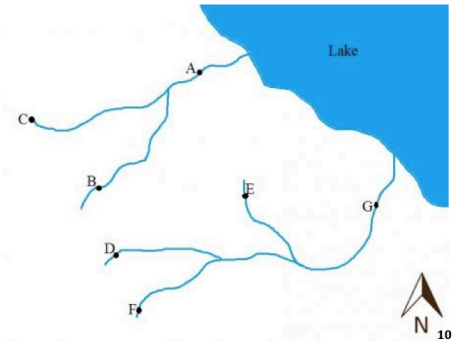
12. Where will a lake form in the topographic map above?

- A
- B
- C

13. Where is a hill located on the topographic map above?

- A
- B
- C

For the following items, refer to the image below which represents an overhead view of a watershed



14. Suppose there is pollution at Site D, which location would you expect would become polluted?

- Site F
- Site E
- Site G
- Site B

15. Based on the watershed map, which location would you predict to have the lowest elevation?

- Site D
- Site F
- Site E
- Site G

16. Based on the watershed map, which two sites are located on different watersheds?

- D,G
- B,A
- E,F
- D,B

For the following items, refer to the image below, which highlights the Mississippi River watershed, which is outlined by a dashed line.

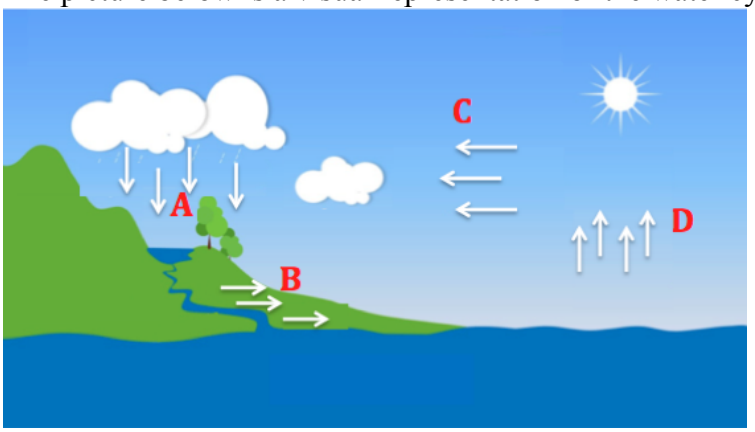
17. The image above highlights the Mississippi river watershed. If the red dots indicate polluted water and the blue dots indicate clean water, where is the most likely source of the pollution?

- A
- B
- C
- D

18. In which state is it possible for the rainfall to end up in the Mississippi River?

- California (CA)
- Washington (WA)
- North Dakota (ND)
- New Jersey (NJ)

The picture below is a visual representation of the water cycle.



19. In the picture above, which group of arrows represents an instance of evaporation?

- Group A
- Group B
- Group C
- Group D

20. Under which conditions in the picture above would you expect the rate of evaporation to be the greatest?

- 15ºF night at midnight
- 30ºF night at midnight
- 70ºF at noon
- 45ºF at noon

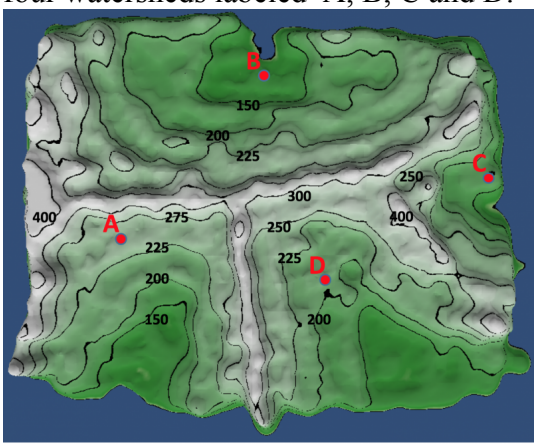
21. Which is an example of evaporation?

- ○ Water vapor becomes liquid water on the side of a glass
- ○ Water vapor moves from inside a cell wall to the atmosphere
- ○ Liquid water moves through soil particles
- ○ Liquid water heats up and turns into water vapor

22. What is evaporation?

- ○ The process by which particles leave a gas and become a liquid
- ○ The process by which a solid changes directly into a gas
- ○ The process by which particles leave a liquid and become a gas
- ○ The process by which a liquid changes into a solid

For the following items, refer to the topographic map below which depicts an island with four watersheds labeled A, B, C and D.



23. If it rains the same amount on the whole island, which watershed would output the most water into the ocean?

- A
- B
- C
- D

24. If it rains the same amount on the whole island, which watershed will have the lowest amount of water flow into the ocean?

- A
- B
- C
- D

Scientific Argumentation Assessment

John and his family were about to leave for a weeklong trip. He filled the dog's water bowl and placed it on the screened-in porch. However, before leaving John and his parents decided to take the dog on the trip, but John left the water bowl on the porch. One week later when John and his family returned, they noticed that the water bowl was empty.

Curious about the empty bowl, John tells both his brother (Matt) and sister (Cathy) about the missing water. Matt and Cathy share their ideas about why the bowl was empty.
Matt says: The neighbor's dog, Rex, drank the water.
Cathy says: The water evaporated into the air.

After hearing both Matt's and Cathy's ideas about why the bowl was empty, John begins to do some research of his own and collects the following pieces of information. Whose idea does each piece of information support?

	Matt	Cathy	Both
1. Hot weather causes water to evaporate more quickly.	○	○	○
2. The screen Porch is undamaged.	○	○	○
3. The water bowl does not have any cracks.	○	○	○
4. The neighbor said that Rex was outside all week.	○	○	○

Matt and Cathy elaborated on their ideas about why the bowl was empty.
Matt says: The neighbor's dog, Rex, drank the water. The bowl is empty now because Rex got thirsty.
Cathy says: The water evaporated into the air. During the trip, the temperatures averaged 98ºF throughout the week. The higher temperatures cause the water to evaporate quickly and by the time we got back home the water had already evaporated.

5. What is Matt's claim about what happened to the water?

- ○ The neighbor's dog drank the water
- ○ The bowl is empty now
- ○ Rex got thirsty

6. What is Cathy's claim about what happened to the water?

- ○ The water evaporated into the air
- ○ During the trip, the temperatures averaged 98ºF throughout the week
- ○ The high temperature during the week speed up the evaporation of the water

7. What evidence does Cathy use to support her claim?

- ○ The water evaporated into the air
- ○ During the trip, the temperatures averaged 98ºF throughout the week.
- ○ The higher temperatures cause the water to evaporate quickly and by the time we got back from the trip the water had already evaporated.

8. What reasoning does Cathy use to support her claim?

- ○ The water evaporated into the air
- ○ During the trip, the temperatures averaged 98ºF throughout the week.
- ○ The higher temperatures cause the water to evaporate quickly.

In science class, two students, Oona and Amanda, presented their arguments about what happened to the water.

Oona says: We know the following things: The weather was hot the week the family was gone, the neighbor's dog was outside all week, the water bowl was not cracked, and screen porch was undamaged. Considering all of this evidence, I think the neighbor's dog drank the water in the bowl. Over the week the temperatures averaged 98ºF. When the dog got thirsty, he drank the water in the bowl.

Amanda says: We have the following evidence: The weather was hot while the family was gone, the screen porch was undamaged, the water bowl was not cracked, and the neighbor's dog was outside all week. Based on this evidence, I think that the water went into the air. Over the last week the temperatures were very hot. The hot weather caused all of the water in the bowl to evaporate and there was none left at the end of the week.

9. Between Oona and Amanda, whose argument is better?

- ○ Oona's Argument
- ○ Amanda's Argument

10. Both Oona and Amanda presented their arguments to their science teacher, and the teacher identified Amanda's argument as better. Why do you think the teacher thought Amanda's argument was better?

- ○ Amanda used more evidence in her argument than Oona
- ○ Amanda considered all of the important pieces of evidence. Oona did not consider an important piece of evidence.
- ○ Amanda's argument includes evidence that supports her claim. Oona's argument does not include evidence that supports her claim.

The next day in class, two other students Emily and Juan discuss what happened to the water.
Emily says: The water evaporated into the air. During the trip, the temperatures averaged 98ºF throughout the week. The higher temperatures caused the water to evaporate quickly and by the time we got back home the water had already evaporated.

Juan says: The water went into the air. I know that water evaporates all the time, so if you leave a bowl of water outside for a long time it will eventually evaporate.

11. Between Emily and John, Whose argument is better?

- ○ Emily's argument
- ○ Juan's argument

12.Emily and Juan presented their arguments to their science teacher, and the teacher identified Emily's argument as better. Why do you think the teacher thought Emily's argument was better?

- ○ Emily's argument has evidence. Juan's argument does not have evidence.
- ○ Emily's argument contains reasoning. Juan's argument does not include reasoning.
- ○ Emily's claim is consistent with her evidence and reasoning. Juan's claim is not consistent with his evidence and reasoning.

VITA

Eric Wulff was born in Chicago, Illinois, but spent most of his childhood in Lansing, Michigan. Eric graduated from Michigan State University with a Bachelor of Science degree in Zoology in 2010. Following graduation, Eric then began a graduate program in Biology at West Virginia University, working in the area of reproductive physiology. After working for a year in a Biology laboratory, Eric then deiced to pursue a career in education and graduated with a Master of Education degree in Learning, Teaching, and Curriculum at the University of Missouri. Eric then taught secondary Biology for three years in Eldon, Missouri. After this time, Eric then transitioned back to the University of Missouri to pursue a Ph.D. in Learning, Teaching, and Curriculum.