Summer 2017

Argument education in higher education: A validation study

Paul E. Mabrey III
James Madison University

Argument Education in Higher Education: A Validation Study

Paul E. Mabrey III

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

School of Strategic Leadership Studies

August 2017

FACULTY COMMITTEE:

Committee Chair: Dr. T. Dary Erwin

Committee Members/ Readers:

Dr. Karen Ford

Dr. John D. Hathcoat

Dedication

For Erika and Gene Mabrey, thank you for all of your help, encouragement, collaboration, and motivation. I could not have done with without your support, sacrifices, inspiration, and love.

Acknowledgements

The journey to and through this dissertation would not have been possible without the community of family, friends, colleagues, faculty, students, colleagues, debaters, coaches, and so many more who have supported me. Thank you. A special thank you to the members of my dissertation committee; Dr. Dary Erwin, Dr. Karen Ford, and Dr. John Hathcoat. I am grateful the opportunity to have worked with Dr. Erwin, who helped guide me through the doctoral program, mentored me on postsecondary leadership, and motivated me to be interested in higher education policy.

Table of Contents

List of Tables

Abstract

Argument education can play an important role in higher education for leadership development and responding to increasing calls for post-secondary accountability. But to do so, argumentation teachers, scholars, and practitioners need to develop a clearer definition and research agenda for the purposes of teaching and assessing argumentation. The research conducted here contributes to this project by first establishing a definitional construct and observable behaviors associated with learning and practicing argumentation. Second, an argument education assessment instrument was created based off of the literature-supported definition of argumentation. Third, debate and argument education subject matter experts reviewed the definition, behaviors, and assessment instrument. Fourth, the newly developed instrument was administered to undergraduate college students over the course of three studies (n=949) to collect evidence testing whether the instrument may be used in a reliable and valid way to assess the learning of argumentation. Finally, the author concluded that the data suggests that the instrument may be used for assessing argument education, but further research is needed to improve the evidence for reliability and validity of the instrument's use. Furthermore, the data collected from assessing argument education provides important implications for how argumentation is defined and assessed within an educational context and what role argument education may play in leadership development.

CHAPTER 1

Introduction

Argumentation may be overlooked due to negative connotations within public discourse, individual experiences with interpersonal conflict, and/or a general lack of familiarity with the term. Such issues may partly be attributed to inadvertently conflating argumentation with other skills, such as critical thinking and/or problem solving (Paris, 2016). But despite this, argumentation remains a foundational discipline and educational approach that dates back to the Ancient Greeks. Argumentation, according to van Rijn, Graf, & Deane (2014), "is not only important in the language arts, but also in mathematics and science" (p. 110). For many disciplines like history, mathematics, and science – argument is an essential skill set to academic and professional success. Students need to be able to evaluate evidence, develop interpretations, analyze the arguments of others, and make their own case. And this skill set is not restricted to academia but transfers well to outside audiences. According to Osborne (2010), "What is in little doubt is that employers, policymakers, and educators believe that individuals' ability to undertake critical, collaborative argumentation is an essential skill required by future societies (47)" (p. 466). The skills may not always be perceived or labeled as argumentation but the underlying construct and observed behaviors are based in argument. And the learning and practice of argumentation may be a great benefit to those within and beyond postsecondary education.

Argument Education as Leadership Development

One benefit to argument education across higher education is preparing and developing future leaders. Leadership development, sometimes referred to as leadership

education, is the practice that explicitly aims to provide training opportunities for potential (or current) leaders to develop productive leadership behaviors, styles and characteristics. Day (2012) advances this understanding by claiming that "[t]he notion of roles and processes refers to behaviors or other actions enacted by anyone – regardless of whether or not considered as a formal leader – that facilitate setting direction, creating alignment, and building commitment" (p. 108). Here, leadership development presumes that one can be taught to embody or practice the essential elements of leadership. Leadership development is aligned with a framework that presumes leadership behaviors are malleable rather than natural or evolved. This approach then supports the notion that leadership interventions are not only possible but can be effective.

Argument education has the potential to improve student learning and application of essential leadership concepts, like reasoning, decision-making processes and empathy. In particular, these skills can help students develop toward being transformational leaders. Antanokis (2012) reviews transformational leadership and characterizes it as concerned with the leader-follower interaction. Transformational leadership includes aspects of the softer side of leadership like vision, motivation and charisma while simultaneously being concerned with accomplishing the tasks required of a given situation. Transformational leaders are required to know their followers, audience, and situation and then build a persuasive case toward some visionary path or action. It might be called idealized influence, intellectual stimulation, individualized consideration, or contingent reward, but what all of these characteristics have in common is the ability for a leader to identify a situation that requires influence and develop the arguments appropriate to motivate within that environment.

Curricular interventions centered on argument education are uniquely situated to help develop leadership in postsecondary students. Argumentation can actively engage students though simulation, role-playing and actual debates. Through argument-based pedagogy, students are asked to practice evidence-based decision-making from different perspectives and in a variety of contexts. Student teams are asked to research interesting contemporary topics while developing and communicating controversial positions. These kinds of activities, according to Rao (2010), "[p]rovide for individual construction of holistic knowledge in a collaborative atmosphere lending itself to an engaging learning experience" (246). Throughout the process of debates or role-playing, students are required to actively listen and understand the position of others in order to be successful. While the potential exists for debate to impact leadership development, little research has been done. This project is an attempt to explore if various argument education approaches are effective interventions for increasing argumentation skills.

Argumentation Across the Curriculum

With the successes of urban debate leagues in middle and high school, we are witnessing more attempts to integrate argument-based education into the curriculum nationwide (Deards, 2014). For example, Yanklowitz (2013) wrote "Critical thinking and dialogue are often made manifest in the form of argument." He goes on to suggest that training in argument is one of the best ways to improve critical thinking skills and that our education systems should do more to integrate this into our school systems.

In fact, according to Argument Centered Education (no date), the recent Common Core standards have integrated argument throughout the standards. They claim,

> Argument is the core of the Common Core. Education writers such as Mike Schmoker and Deanna Kuhn have made this point, but the authors of the standards reveal it themselves. Argument is 'the soul of an education,' says the CCSS Research Appendix, because when students are engaged in argument about an issue of importance, 'something far beyond the surface knowledge is required: students must think critically and deeply, assess the validity of their own thinking, and anticipate counterclaims.' College is, they quote Gerald Graff, an 'argument culture,' rigorous college preparation demands first and foremost that students are taught 'argument literacy.'"

While Common Core reflects standards and trends throughout K-12 education, this is still relevant for higher education. Argument culture and argument literacy are important because argument is woven throughout our education, jobs, and civic life. Any attempt to persuade, advocate, or even just convince a friend is based on argument. And yet despite the seemingly overwhelming support for argument as a value in both K-12 and postsecondary education, very few college classes or majors integrate argument education into their curriculum. Debates scaffolded on argument education should be extended throughout the collegiate curriculum, not just practiced in middle and high schools. Llano (2015) claimed that "Ultimately, we could see debating on most campuses helping keep the habit and practice of critical thinking alive not just in select classrooms but as part of what makes the campus experience as intellectually challenging as it is special" (p. 150). But even where there is an argumentation class or argument-based activities, implementing argument education alone is not enough. One must have a plan for assessing the learning of argumentation or argument education. Essential to this

evaluation plan is an instrument that produces results that are valid, reliable, and accessible for their situation.

Statement of the Problem

Institutions of higher education have been facing increasing demands for accountability in two important ways. First, they are asked to justify the value of a college degree (Leonhardt, 2014). Second, stakeholders of the college community are being asked to provide more substantive and data-driven responses to the calls for accountability. One only need to look at the headlines of major newspapers, education industry journals, policy think tanks or public opinion polling to see overwhelming evidence of these growing demands (Hamilton, 2010; Stratford, 2015).

To answer the first question about the value of a college degree, respondents have long replied with a variety of skill sets or behaviors acquired through a college education (Christie, 2014; Cook, 2015). For example, critical thinking, communication and interpersonal skills have been claimed as the value added benefits to obtaining a degree (Berrett, 2013; Davidson, 2016; Gallo, 2014; Iowa State University, 2016). The second question is being answered with more assessment, research and data collection regarding college participation versus not participating in some form of postsecondary education. These efforts may take the form of various classroom and out of class activities such as research about teaching and learning in the classroom, student affairs programming, counseling best practices, advising, and alumni engagement surveys. What is needed is the identification of particular interventions and high-impact practices. This research is an attempt to answer both of those calls with a study to explore if various argument education approaches are effective interventions for increasing argumentation skills.

This study contributes to argumentation education and assessment in higher education by reviewing the literature on debate and argument education in chapter two. I review the method for designing and testing an argumentation assessment instrument in chapter three. In chapter four I present the results of the research studies and discuss the implications of these results for argument education across higher education in chapter five.

CHAPTER 2

Literature Review

Defining Argumentation

Scholars have approached argumentation from different approaches (Andrews, 2009a; Deane & Song, 2015; Zarefsky, 2001; Zarefsky, 2014). For example, Deane & Song (2015) represent a more rule bound approach to argumentation for they describe argumentation "as a kind of *dialectic* – a rule-governed form of discussion in which various speech acts (including assertions, questions, and explanations) are coordinated in the service of social norms for collaborative reasoning (van Emeren & Grootendorst, 1992)" (p.3). In contrast, Andrews (2009a) offers argumentation as "the process of developing arguments, the exchange of views, the seeking and provision of good evidence to support claims and propositions – the *choreography* of argument" (p. 39). Andrews draws attention to argument as an art rather than a strict rule governed technical exchange. Missing from Andrews' definition is argument to what end or for what purpose. In Deane & Song, argument is coordinated toward the social norms of reasoning together. Zarefsky (2001) offers a definition encompassing elements of both definitions. He describes argumentation as

> [T]he study of reason-giving used by people to justify their beliefs and values and to influence the thought and action of others. Its central concern is with the rationality or reasonableness of claims put forward in discourse. This, in turn, depends on whether the claims are warranted, or grounded in evidence and inference that are themselves acceptable and hence constitute good reasons for the claim. (p. 33)

Here, Zarefsky provides a goal of argument, to influence the thought and actions of others, in addition to how argument happens. Rationality is the main tenant of argument for him and arguments must flow reasonably. For arguments to take place, they generally need to include some claims, warrants, and evidence all connected with one another. And even though argumentation must be rational, it must be rational within the realm of influencing others, necessitating the considerations of one's audience or situation.

Teaching Argumentation

Zarefsky's definition of argumentation, or ones like it, has been used to teach argument across educational settings and disciplines. For example, scholars have studied the use of argumentation within history, science (elementary school and post-secondary), and calculus (Andrews, 2009b; Bathgate, Crowell, Schunn, Cannady, & Dorph, 2015; Kwon, Bae, & Oh, 2015; Osborne, 2010). In each of these disciplines, the use of argumentation to teach students was important because subject matter itself, the authors argued, either was or required argument. History and science were each defined as a series of arguments while calculus required students to make arguments throughout their mathematical proofs.

Argumentation research studies also sought out to identify some of the essential skill sets for argument. The study conducted by Bathgate, et al (2015), most resembled Zarefsky's approach to argumentation. They identified key skills relating back to scientific argumentation; like evaluating evidence, justifying argument, and understanding that the social context of different perspectives is important. These skills align closely with Zarefsky's emphasis on rational arguments delivered to influence a particular audience. Deane and Song (2015) also propose a framework of learning

progressions for teaching argumentation across different developmental levels. Argumentation, according to them, included five phases. They are understanding the issue (appeal building), exploring the subject (inquiry and research), considering the positions (taking a position), creating and evaluating arguments (reasons and evidence), and organizing and presenting arguments (framing a case). These five phases also closely resemble important aspects of the Zarefsky definition of argumentation.

Researching Argumentation

While scholars have conducted studies that have identified positive benefits to using argumentation as a teaching instrument, more research is still needed (Andrews, 2009b; Bathgate, et al, 2015; Deane & Song, 2015; Hasnunidah, Susilo, Irawati, & Sutomo, 2015; Kwon, Bae, & Oh, 2015; Leite, Mouraz, Trindade, Martins Ferreira, Faustino, & Villate, 2011; Osborne, 2010; vin Rijn, Graf, & Deane; 2014). These studies have used a variety of methodological approaches to demonstrate the impact of student exposure to argumentation training. They have utilized close textual analysis, student interviews, subject matter tests, willingness to argue scale, multiple choice (with two open-ended questions) instrument, writing assignments, and classroom journals. For example, Hasnunidah et al (2015) relied on writing assignments to solicit samples that could be evaluated for argumentation. Hasnunidah et al used integrated writing prompts to measure argumentation and critical thinking through pre/post tests. The argument rubric used here was based on Toulmin's model of claim, data, and warrant. They found that students exposed to scaffolded argument interventions in a biology class scored higher on these essays rated for argument and critical thinking than students in a standard lecture biology course. Hasnunidah et al concluded, "The improvement of the

argumentation quality might affect to the improvement of the critical thinking skill of the students" (Hasnunidah et al, 2015, p. 1191). But even with this research, scholars have suggested that more empirical research on argument education is necessary. For example, Bathgate, et al (2015) asked "But empirical evidence on the benefits of argumentation ability for science learning is still lacking; do students with such abilities actually learn more science content than students who do not have such abilities?" (p. 1592). Or Osborne (2010), for instance, suggested that "Research on the development of students' skills in argumentation is still in its infancy and lacking valid or reliable instruments with which students' competency can readily be assessed" (p. 466). Even though preliminary evidence has been gathered to demonstrate the positive impact argumentation can have on student learning, more rigorous empirical studies are needed.

Defining Debate

Debate, in some form or another, has long been part of social, academic, and political life. In fact, Vo and Morris (2006) claim it is common knowledge that "debating as a teaching tool has an honorable tradition" (p. 315). Dating back to the days of Aristotle and Plato, debate has been used as a method for teaching content, skills, attitudes, ethics, civic life, and more. And this is still true today. Debate is used curricularly and extra-curricularly to teach knowledge, skills, and attitudes across, in disciplines, like business, dentistry, accounting, economics, communication, and technology studies, social work, biology, health care, medical school, environmental science, and computer science (Camp & Schnader, 2010; Darby, 2006; Goodwin, 2003; Gregory & Holloway, 2005; Jagger, 2013; Jerome & Algarra, 2005; Koklanaris, MacKenzie, Fino, Arslan, & Seubert, 2008; Lilly, 2012; Nguyen & Hirsch, 2011; Proulx,

2004; Rao, 2010; Roy & Macchiette, 2005; Scott, 2008; Vo & Morris, 2006; Winkler, 2011).

Debate as a teaching tool shares several important common characteristics, regardless of the context or discipline. For example, Roy and Macchiette (2005) propose some basic guidelines for utilizing debate in the classroom. Debates should involve students giving oral arguments, supported by researched evidence, for or against a controversial topic. Sometimes these debates have students one on one, two on two, or in some other format. The controversy to be debated typically depends on the class content. If the class is interdisciplinary or skill based, the topical content may be generated by student interest. The students debating should conduct the research themselves, relying mostly on scholarly sources. Based on the research, students generate arguments in response to the assigned topic. Students outside of class generally conduct all of the work conducted to this point, often times collaboratively with members of their team or group. The students then carry out the actual format of the debate within class, students speaking directly to and in front of one another on the topic at hand. They make arguments, respond to the arguments of other students, synthesize content, evaluate evidence, make summary judgments about the controversy, etc. At the conclusion of a given classroom debate activity, the debating students are typically given feedback by the faculty member and sometimes their student peers sitting in the audience. These evaluations are often based on content knowledge, refutation skills, quality of research, public speaking delivery, or other criteria determined by the faculty and/or class.

Debate as a Teaching Tool

Many positive benefits to debate participation have been identified because of the format and elements involved in debating (Darby, 2006, Goodwin 2003). For example, students and faculty have found debate to help improving content mastery, addressing controversial topics, developing communication skills, improving critical thinking, decreasing discipline referrals, argumentation skill confidence, and bettering research practices (Camp & Schnader, 2010; Darby, 2006; Gregory & Holloway, 2005, Goodwin, 2003; Rao, 2010; Roy & Macchiette, 2005; Scott, 2008; Vo & Morris, 2006; Winkler, 2011). For example, Camp & Schnader (2010) suggested that "Debate encourages students to develop research and presentation skills, apply their knowledge in a logically consistent manner, and interact with peers in a meaningful way" (p. 658). This occurs in part because of the active learning required from debate assignments, and not as what Darby (2006) referred to as "a test of knowledge acquired" (p. 2). Among all of the learning benefits, critical thinking is probably the benefit most often cited from student debate participation.

Using Debate to Teach Critical Thinking

Critical thinking is a student learning outcome often cited from debate participation in the classroom (Berkowitz, 2006; Camp & Schnader, 2010; Jackson, 1973; Llano, 2015; Nguyen & Hirsch, 2011; Rao, 2010; Roy & Macchiette; 2005; Scott, 2008; Tous, Tahriri, & Haghighi, 2015; Vo & Morris, 2006). Several of the studies cite Facione's definition of critical thinking as a starting place. For example, Berkowitz (2006) refer to Facione's "summarizing the results of a consensus of experts, indicated that the core cognitive skills of critical thinking are interpretation, analysis, evaluation, and inference" (p. 45). Roy & Marcchiette (2005) build a theoretical case for how debate

fosters critical thinking in students throughout the classrooms. They say, “Critical thinking allows students to reach beyond a single perspective, to challenge assumptions, and to better analyze a wide range of challenges and problems in adult life” (p. 265). Students are able to learn these skills because of the debate format that encourages researching and exploring multiple positions on a given controversy.

Researching Debate as a Teaching Tool for Critical Thinking

Even though debate scholarship cites critical thinking among its education benefits, very few studies are able to empirically observe it. Two studies sought to demonstrate improved critical thinking by having students complete a 10-item self-assessment after participating in their class debate assignments (Rao, 2010; Roy & Marcchiette, 2005). Faculty and students reported that the debate assignments did increase student critical thinking, but Vo & Morris (2006) claimed that, “we are not sure that we have seen empirical works specifically designed to measure learning outcomes of debate used as a supplementary tool” (p. 319). Camp & Schnader (2010) conducted a different study utilizing a pre/post survey, self-assessment, and a free response specifically about critical thinking. And while they found evidence of a positive impact on critical thinking from debate participation, their study also relied on more indirect measures like student self-reports. A meta-analysis conducted by Berkowitz (2006) reviewed some empirical support for the impact of debate participation on critical thinking, but it is a much more general approach. She combines public speaking, argumentation, debate, and forensics interventions throughout her analysis. And even then she treats all types of interventions as the same within each category. The California Critical Thinking Skills Test (CCTST) and Watson-Glazer Critical Thinking Test were

cited among studies that did attempt to gather some empirical evidence to support their claim that debate pedagogy can positively impact critical thinking skills (Berkowitz, 2006; Tous, Tahriri, & Haghighi, 2015). But even though these forms of more direct evidence are preferable, their studies seemed to be in very unique circumstances and are not as generalizable. For example, the Tous, Tahriri, & Haghighi study used the CCTST to explore how debate as a teaching tool might impact the relationship between reading comprehension and critical thinking among 120 Iranian high school students. Given the strong belief that debate as a teaching tool can impact student critical thinking but lack of more direct empirical evidence, it may be beneficial to turn to another field to better understand how debate impacts critical thinking and whether or not this has been observed through research.

Using Debate to Teach Argumentation

Argumentation studies is a good discipline to supplement debate literature because debate practitioners develop their practices, assignments, and debate teaching tools based on approaches to argumentation. Questions of what counts as evidence, what makes an argument, how does one engage in argument, applying arguments to a given context, or even how to craft a controversial topic have their root in argumentation studies. Argumentation as an academic discipline has a richer history to draw from than does the literature on debate pedagogy or practice. Argumentation studies can provide insight into how to define argumentation and how argumentation has been taught and studied. Debate can also look to argumentation research to better understand, explain, and even further research the relationship between debate and critical thinking.

Some argumentation scholars have made the link between argumentation and critical thinking (Andrews, 1995; Hasnunidah et al, 2015). They suggest that scholars and practitioners should focus on teaching and studying argumentation rather than critical thinking. Richards (1995) explains this when he makes three reasons for focusing on argument rather than critical thinking. First, "[A]rgumentation is social, dialogic (or multi-voiced), and tangible. You can see evidence of it, and therefore subject it to critical analysis" (p. 42). Second, Richards claims that argument enables feelings, emotion, and affect to be considered whereas critical thinking is perceived as a focus away from feelings. Third, argument is more attentive to context while critical thinking is concerned with process and procedure. The empirical study conducted by Hasnunidah et al (2015) goes one step further suggesting that argumentation is related to and actually a precursor to critical thinking. Their study utilized an essay test to measure the argumentation and critical thinking skills of 180 pre-service science teachers. In it, they found that the

> Lower argumentation skills of students into one of the causes of low student critical thinking skills. The fact of the results of the survey showed that the critical thinking skills of is still low. This is evident from several indicators, among them: students have difficulty in asking the questions and defining the problem, the literacy of the actual problem is still lacking, problem solving analytical and evaluative biology is still low, skills to identify, analyze, and evaluate arguments selectively is still low. (1186).

Throughout their study, they found that essential argumentation skills like exploring the multiple issues within a controversy, building a case for one position while also

understanding the context of other perspectives, and judging the quality of evidence or an argument are fundamental to the development of strong critical thinking skills.

Debate is uniquely capable of teaching argumentation as a pedagogical approach. Debate, more than other types of classroom assignments or approaches to teaching, is able to tap into the social and dialogic aspect of argumentation. In asking students to participate in a classroom debate, faculty are situating the student within a risky social context. Students have to orally articulate a position with well-supported arguments in front of and alongside their peers. In preparation for this debate, students must engage in research exploring the multitude of perspectives that surround a given controversy. Furthermore, students are asked to anticipate the arguments that their debate opponents or different stakeholders (depending on debate format) may take during the debate assignment. And what is perhaps the most daunting ask of students (and most unique to debate), the positions created and articulated are challenged on the spot and a given student will be asked to respond and defend their argument or position in the moment. As Deane and Song (2015) suggest, one of the best ways to develop argumentation skills is to create an interactive situation and social requirement for effective argument. In this way, classroom debates offer a potentially invaluable teaching tool for developing argumentation skills in students throughout higher education. But missing in this conversation is how should one go about measuring if debate interventions are successful in helping students learn argumentation.

Review of Existing Argumentation Assessment Instruments

Attempts to assess argumentation and/or debate interventions typically address one or more of the following: satisfaction with the activity, agreement with the topic,

content or knowledge improvement, critical thinking, and argument. And what has been reported in the literature often does not share very much information about the instruments. This reflects a fairly underdeveloped, or at least unpublished, approach to assessing and measuring argument education in higher education.

Satisfaction is probably the most often used assessment for argumentation and debate classroom activities (Goodwin, 2003; Gregory & Holloway, 2005; Koklanaris et al, 2008; Rao, 2010; Vo & Morris, 2006). In these studies, students are administered some form of survey either right after the debate activity or toward the end of the semester. The instruments are typically some form of likert-scale items and may include open-ended questions. For example, Rao and Vo & Morris used a ten-item self-reported satisfaction and learning instrument. They reported a Cronbach's alpha of 0.88. Goodwin, on the other hand, used informal classroom discussion and open-ended written responses to collect student satisfaction with the debate activities.

Topic agreement is another form of assessment used, though usually associated with controversial topics (Lilly ,2012). Here, faculty will administer a likert-scale based survey soliciting students' opinions about a topic. They do this before a debate activity and afterwards to gauge if students have changed their opinions about a topic as a result of participating in the debates. Lilly, for instance, asked students if they agree or disagreed with the position they debated in their college environmental science course. The question here was a simple yes/no survey given before and after the debate.

Assessing the impact of debates on content knowledge is another common way to measure the impact of using argument or debate activities in the classroom (Camp & Schnader, 2010; Koklanaris et al, 2008). These forms of assessment are generally pre-

and post-tests created by the faculty of that specific course to measure any difference in content knowledge as a result of the intervention. The tests may be administered within a class that has integrated an intervention or between classes that used different interventions (or no intervention). These instruments are usually unique to a given faculty member and their class because of the associated learning outcome of increasing course knowledge. Koklanaris, for example, developed a 10-question multiple-choice health sciences quiz that was administered before and after an intervention. One group used debates in the class while the other group attended traditional lectures.

Critical thinking is another construct or learning outcome that is assessed alongside argumentation and debate interventions (Berkowitz, 2006; Tous et al, 2015). This area of learning assessment is perhaps the most developed, when used, because it utilizes instruments from a more mature assessment and measurement field. More developed because the critical thinking assessment tools used are often commercial instruments that have been well developed and validated. Berkowitz, in her meta-analysis, reviewed 23 studies that attempted to measure the impact of debate, forensics, and public speaking on critical thinking. In the review, she found that while several of the difference commercial instruments were used, the Watson-Glaser Critical Thinking Appraisal was the instrument used most often. While the instruments are more developed, they are not used very often because of the costs associated with using them.

Very little research has been conducted to indirectly or directly measure argumentation. More often, argumentation is studies as a vehicle to impact other constructs, skills, and observed behaviors like the ones mentioned previously. Of those measuring argumentation, indirect and self-reported measures are the more common

attempts. Gregory & Holloway (2005) used a pre/post confidence in oral and written argument survey the administrated to assess the impact of classroom debate participation on argumentation skills. No test information was reported about their survey. Again, this is a place where some of the satisfaction post-surveys are used though they are not a very developed area for assessing actual argumentation skills. While not as prevalent, a few scholars have attempted to more directly measure argumentation skills through rubrics and constructed response (Bathgate et al, 2015; Hasnunidah et al, 2015). The study by Hasnunidah and colleagues used an analytical framework based off of Toulmin's model of argument (claim, data, warrant) to rate pre/post essays. The framework provided a scoring range from 1-5 on the singular framework. The study did not provide the prompt for the study, but did report a reliability index of 0.690 for the argumentation test. But it was not clear from the study which estimate of inter-rater reliability was used. Nor did the study go into details about the raters or rating process. The Bathgate et al study provided the richest and most rigorous example of an instrument designed to measure argument. In their study, they were concerned with scientific argumentative sense making for middle school students. They contextualized this into two different parts, justifying argument and anticipating the arguments of an opponent. The researchers developed a nine-item instrument that included seven multiple-choice items and two open-ended items. The measure was created in consultation with a discipline context expert. A coding criterion was created for each of the two parts; 0-5 for argument justification and 0-4 for anticipating others' arguments. The authors did report Cohen's Kappa ranging from 0.87 to 0.93.

A review of the literature on argumentation and debate education reveals no shortage of attempts to implement debate activities into the classroom as a way to develop argumentation skills. Argument-based interventions are happening in disciplines across the curriculum, though the research on these curricular innovations seems to be more concerned with sharing of ideas and programming rather than demonstrating that learning is happening. Ample singular anecdotes exist that speak to the potential for argument education to add value to a student's learning during college; providing the knowledge, skills, and attitudes necessary to be a transformational leader. However, what is missing is a research agenda that attempts to measure and demonstrate that the learning and practice of argumentation is happening linked to these curricular argument-based interventions. The research conducted here hopes to contribute to these literature bases by making a call for more empirical research on the impact of argumentation education. The evidence for argument's impact needs to move beyond self-reported and indirect measures of learning to more direct, observable, and replicable studies. But more than just a call, this study begins the process of developing and validating an argument education instrument that can be used for teaching, learning, and study argumentation across higher education.

Research Hypotheses

I propose the following four hypotheses to study and better understand argumentation education:

Hypothesis 1. The argumentation education assessment instrument will yield a generalizability-coefficient greater than 0.70.

Hypothesis 2a. Students participating in a collegiate curricular intervention where debate pedagogy has been integrated into the curriculum will score higher on an argumentation education instrument than students in a control group.

Hypothesis 2b. Students participating in a collegiate extra-curricular debate intervention will score higher on an argumentation education instrument than students in a control group and students in the collegiate curricular intervention.

Hypothesis 3a. Students participating in a collegiate curricular intervention where debate pedagogy has been integrated into the curriculum will score higher on an argumentation education instrument than before their intervention.

Hypothesis 3b. Students participating in a collegiate extra-curricular debate intervention will score higher on an argumentation education instrument than before their intervention.

CHAPTER 3

Method

Measure

Instrument development

The argumentation assessment instrument was developed by the author in consultation with a higher education policy and assessment subject matter expert. The product and process were both influenced by the Ennis-Weir Critical Thinking Essay Test (Ennis & Weir, 1985) and the National Assessment of College Student Learning, conducted by the National Center on Postsecondary Teaching, Learning, and Assessment (Jones, 1995).

The process began by reviewing the Jones (1995) section on critical thinking. From the lists of identified and agreed upon essential critical thinking behaviors, the author selected the ones consistent with the definition and observed skills for argumentation. From the list of observed argument as critical thinking behaviors, the author abstracted out the larger skill set or component of argumentation. For example, see Table 1 for the list of observed behaviors from Jones (1995) and how this author has grouped them into argument skill set themes. These 21 behaviors formed 5 different skills from within argumentation. The five skill sets are:

- Identify biased argument (3 behaviors, for example “Recognize use of misleading language”)
- Prioritize information based on the situation (5 behaviors, for example “Detect introduction of irrelevant information into an argument”)

- Argument construction (5 behaviors, for example “Determine if one has sufficient evidence to form a conclusion”)
- Argument evaluation (6 behaviors, “Evaluate an argument in terms of its reasonability and practicality”)
- Argument utilization in a situation (2 behaviors, for example “Present supporting reasons and evidence for their conclusion(s) which address the concerns of the audience).

Each one of these skills then became an item on a rubric. Identify affective argument was added as a sixth skill because the social, interactive element is a critical element of argumentation but not explicitly present in the Jones work. Affective argument is operationalized here as the emotions, feelings, attitudes, values, or other relational dimensions that play an important role in argumentation. The rubric was used to score the written responses solicited from respondents via short answer prompts. The prompts were designed such that respondents were demonstrating competency in these different argumentation areas. These are behaviors raters identified when reviewing the written answers from the student participants.

Table 1
Argumentation Construct Development

Identify biased argument
- Recognize use of misleading language
- Recognize use of slanted definitions/comparisons
- Determine if an argument rests on false, biased or doubtful assumptions
Prioritize information based on situation
- Detect introduction of irrelevant information into an argument
-Recognize relationship between communication purpose and ideas that must be resolved to achieve this purpose
- Identify background information provided to explain reasons which support a conclusion
- Assess the importance of an argument and determine if it merits attention

- Judge what background information would be useful to have when attempting to develop a persuasive argument in support of one's opinion

Argument construction
- Identify the unstated assumptions of an argument
- Determine if one has sufficient evidence to form a conclusion
- Present an argument succinctly in such a way as to convey the crucial point of an issue
- Cite relevant evidence and experiences to support their position
- Seek various independent sources of evidence, rather than a single source of evidence, to provide support for a conclusion

Argument evaluation
- Evaluate an argument in terms of its reasonability and practicality
- Evaluate the credibility, accuracy and reliability of sources of information
- Assess statistical information used as evidence to support an argument
- Assess how well an argument anticipates possible objections, offers, when appropriate, alternative positions
-Determine and evaluate the strength of an analogy used to warrant a claim or conclusion
-Determine if conclusions based on empirical observations were derived from a sufficiently large and representative sample

Argumentation – argument utilization in a situation
- Present supporting reasons and evidence for their conclusion(s) which address the concerns of the audience
- Develop and use criteria for making judgments that are reliable, intellectually strong and relevant to the situation at hand

---

After identifying these skill sets or sub scales to argument as critical thinking, the next step was to create the rubric (Appendix A). For this, a three category gradient scale (unsatisfactory, fair, or good) with weighting of 0, 1, or 2 respectively was used for each of the six items making up the subconcepts of argumentation. More specifically, zero communicates the lack of evidence for this particular skill while 2 indicates the presence of the skill at the highest level. The 3-point scale was designed based off of the original behaviors that exemplify argumentation. The main distinction between a 1 and a 2 gradient on the scale was understanding the item but not correctly identifying the item in the prompt. For example in the identifying biased argument item, if a participant

communicated that there was bias within the prompt, they would receive a 1. If they correctly identified the source of the bias, that would earn the participant a 2. But if a participant knew there was bias but mis-identified the source of bias, they would still receive a 1.

The final aspect of initial argumentation assessment instrument construction was creating the prompts. Each prompt was created with the intent of eliciting a response from the participant that could then be rated to determine if or at what level a given argumentation skill was present. In creating each prompt, the author attempted to create scenarios that were as accessible as possible. Accessibility here means minimizing as much as possible the amount of background information or disciplinary knowledge necessary to respond to the prompt. This helps reduce any potential construct-irrelevant variance and puts the focus on the specific argumentation construct being assessed. Two possible prompts were created for each of the six argumentation skill sets. Each prompt was drafted in an attempt to be aligned with the skill and observable behaviors for that argumentative skill set. For these studies of the argument assessment instrument, however, only one prompt for each argument skill set was included (Appendix B).

Reliability

Reliability is an essential component of instrument development. Within the context of educational assessment, reliability is the "consistency of examinees' scores across such facets as occasions, tasks, and raters. In other words, reliability addresses whether an examinee's score would be the same if she were to take the exam on a different occasion, complete different tasks, or be scored by different raters" (Johnson, Penny, & Gordon, 2009, p. 22)." And these concerns for reliable scores are even more

important in testing situations that are perceived as less objective. For example, reliability receives more attention in traditional performance assessments like essay responses or oral presentations than a Likert-scale based instrument. Assessment scenarios where human raters are assigning scores, rather than computers, come under even more scrutiny because of the emphasis placed on human judgment to subjectively assigning scores. Within these types of performance assessment, "Interrater reliability refers to the level of agreement between a particular set of judges on a particular instrument at a particular time. Thus, interrater reliability refers to the testing situation, and not of the instrument itself" (Stemler, 2004, p. 1). In constructing an instrument where raters are assigning scores to evaluate a participant's observed argumentation skills, interrater reliability must be estimated.

Even though reliability is widely understood within measurement and assessment, according to Stemler (2004), interrater reliability has been often misunderstood because it is described as a monolithic concept. Stemler argues that "[T]he widespread practice of describing interrater reliability as a single, universal concept is at best imprecise, and at worst potentially misleading. Instead, researchers and practitioners should begin to use more precise language to indicate the specific type of interrater reliability being discussed" (Stemler, 2004, p. 1). He provides three general categories by which interrater reliability can be described; consensus estimates, consistency estimates, and measurement estimates. Consensus estimates are the most often used and are defined as the percent of agreement among raters. Percent agreement and Cohen's kappa statistic are examples of consensus estimates. Consistency estimates are less concerned with agreement between raters rather than how consistent an observed behavior is rated across

raters. The Pearson correlation is an example of a consistency estimate. Measurement estimates attempt to use all information in a testing situation to determine interrater reliability, not just consensus or consistency. Generalizability theory (g-theory), and its g coefficient, is an example of a measurement estimate because it allows for each component of the testing situation and design to be analyzed. For example, g-theory can parse out variance according to rater, item, persons, occasion, etc.

To estimate reliability for the argumentation assessment instrument, the measurement estimate utilizing g-theory is privileged. G-theory is the appropriate reliability test here because of the ability to isolate multiple sources of error, particularly in a constructed response situation. Alkharusi (2012) claims that g-theory "recognizes multiple sources of measurement error, estimates each source separately, and provides a mechanism for optimizing the reliability." (p. 194). One is able to isolate error due to rater, item, participant, or situation. Within g-theory, these objects of measurement are called facets. They are like variables in other traditional statistical analyses. G-theory also allows one to evaluate the interaction of the different facets, for example one particular rater on one specific item. Furthermore, the statistical test used also allows one to simulate ways to improve the reliability. For example, one can run a decision study (d-study) to determine the impact of varying a facet on the g-coeffecient for reliability. In the d-study, a researcher can increase or decrease the raters or items, for example, to determine how that might impact the g-coeffecient.

Validity

To begin making a case for validity, this study employs an argument-based approach to validity described by Kane (1992). Validity is not something that is

possessed within an instrument across all possible uses, but evidence that must be accumulated for the interpretation of scores in a particular situation. *The Standards for Education and Psychological Testing* define validity "as the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test" (Pitts & Naumenko, 2016, p. 5). Rather than presenting a single piece of evidence to demonstrate whether argumentation assessment is measuring what it claims to measure, one must build an argument for the validity of the instrument. As Kane (1992) elaborated, "It is an 'approach' to validity rather than a type of validity. By emphasizing the importance of specifying the interpretative arguments, this terminology highlights the importance of evaluating assumptions, implicit and explicit" (p. 39-40). Put differently, one cannot simply rely on the objective appeals to a type of validity evidence because even though a piece of evidence appears objective, the case for validity still relies on interpretive work (by the author or reader) whether stated or not. For Kane, it is better to put forth the interpretation and build the argument, thus making available all of the claims for questioning.

The interpretive argument here is that the results from the argumentation education assessment instrument can be used to show evidence of whether or not students in higher education institutions are learning foundational argumentation skills. For this argument to be true, several assumptions or inferences are made. First, argumentation is a construct that can be defined, observed, and measured. Second, the rubric created to measure argumentation education reflects the key elements of argumentation. Third, the prompts designed to solicit observed argumentation behaviors align with the rubric. Fourth, the scores generated by the raters on the rubric for identifying argumentation

skills in students of higher education is reliable. Fifth, the argumentation education assessment instrument is able to detect statistical and meaningful differences between groups presumed to have different levels of argumentation education.

Multiple lines of evidence are needed in order to support these inferences. First, validity is demonstrated via the rigorous Delphi study technique employed by Jones (1995). In this study, they surveyed employers, faculty, and policymakers to determine which concrete, observable behaviors were desired for a given skill, critical thinking in this case. The Jones study used a two round procedure to identify moments of agreement between the three groups on the defining behaviors for the skill set. This approach generated consensus on a core set of behaviors that expert stakeholders across different disciplines and industries identified. From these agreed upon behaviors for critical thinking, this author went through and identified the ones that most closely aligned with skills associated with argument as determined by the literature and the author's 15+ years studying and practicing argument.

Second, adding to the evidence for validity, the author will ask subject matter experts in the field of argumentation to review the instrument. These argumentation experts are faculty who study, research, and teach argumentation within postsecondary institutions across the United States. They were sent a survey soliciting feedback on the definition, rubric, prompt, and identified behaviors for argument as critical thinking. The feedback was analyzed as possible evidence for or against the validity of this argumentation education instrument and reviewed for potential future revisions to the instrument. Third, the prompt and rubric were developed and aligned specifically with these identified behaviors. This alignment helps ensure that what is being measured with

the assessment instrument is actually the behaviors associated with argumentation. Finally, the results from the actual instrument can function as validity evidence if the instrument is able to differentiate among the three different sample groups as expected. For example, the extra-curricular group should score more favorably than the control and curricular intervention group because of their more extensive experience practicing and studying argumentation.

Procedure

The author's university Institutional Review Board (IRB) has approved the research protocol for these studies. Four studies were conducted to test the argument education assessment instrument. Study 1 sought out argumentation subject matter experts to review the argumentation construct, rubric, and prompts. Study 2 piloted the instrument as a post-test between three groups (control, curricular, debate extra-curricular) with sample samples. Study 3 expanded on study 2 by increasing the sample sizes, adding an additional curricular intervention group, and administering the instrument as a pre-test to collect longitudinal data. Study 4 replicated the research in study 3.

Study 1

Argumentation subject matter experts via an electronic survey will review the argumentation education instrument (Appendix C). The author identified 19 argumentation experts to send the survey. The survey itself will consist of four major sections, each a mix of Likert-scale and open-ended questions. The first section will solicit feedback on the definition of argumentation informing the instrument. One question will ask if the definition is acceptable on a five-point Likert scale. The second

question will ask if there is anything to include or exclude from the definition. The second section concerns the performance criteria and subsequent expected behaviors. For example, please rate whether you agree or disagree that the following is a foundational argumentation skill, using the five-point Likert scale: "Identify biased argument (recognize use of misleading language, recognize use of slanted definitions/comparison, determine if an argument rests on false, biased, or doubtful assumptions) is a foundational argumentation skill." This section then does this for each of the six skills and then asks if "any of the six should be removed from a foundational understanding of essential argument skills?" The section closes soliciting open-ended feedback about adding any other skills deemed essential.

The third section concerns the construction of the rubric. The subject matter experts are asked if each section of the "performance criteria rubric is clear and reflective of the solicited performance." This is done for each of the six-argumentation skills and is intended to assess the alignment of the rubric to the definition of argumentation and essential observed behaviors. The section closes with an open-ended question about rubric feedback. The final major section for the subject matter experts is to review the scenario prompts. This section asks if the given scenario aligns with the intended argumentation skill. Experts are asked whether they agree on the same five-point Likert scale for each scenario. Again, the section closes with an open-ended question soliciting general feedback about the prompts. The survey closes thanking them for their time and offering an opportunity for any feedback about the argumentation education instrument and overall research project.

Study 2

The argumentation assessment instrument was administered as a pilot study to students in a control group, curricular intervention group, and extra-curricular debate intervention group. The instrument was distributed via electronic survey software to all three groups as a post-test. For the control group, the instrument was included as an option in the research requirement for that course. For the curricular intervention group, the survey was emailed to students in the classes for voluntary completion. For the extra-curricular debate intervention group, the survey instrument was emailed to the Director of Debate at four institutions where the coach agreed to send out to their student debaters. For each of these institutions, their home IRB was contacted and also gave approval as the research involved students at their organization.

For the rating process, the author recruited a faculty member at the author's institution who is a subject matter expert in argumentation to act as one of the two raters for scoring the responses. The author was the other rater. The sample responses were assigned identification numbers after the three sample groups had completed the assessment argumentation instrument. The identification numbers should help ensure that the raters do not know which sample group was represented by the response they were rating. The author conducted initial rater training by introducing the second rater to the research project, rubric, and scenario prompts. The two raters rated the first ten responses for each item separately. After the ten are rated, the raters then discussed how and why each score was assigned. After the first ten responses are rated and the two raters are collaborated, the raters scored the rest of the responses individually with no discussion or agreement. The first ten responses were still utilized in the overall data set. For each of the six items, a response received a 0, 1, or 2 from the raters. The score from the two

raters were averaged to give each participant's response a score for every item (0-2). To calculate an overall total score for the participant, the six item scores were totaled for an overall argumentation score (0-10).

Study 3 and Study 4

The third and fourth study provided the author an opportunity to test the instrument again and add research design layers to enhance the overall study. This study will again use the argumentation education assessment instrument across the three samples (control, curricular, extra-curricular debate) but add two elements. First, a pre-test was added to the post-test. Each group was administered the instrument at the beginning of the semester and then again toward the conclusion of the semester. The same prompts are used for giving the pre- and post-tests. Second, the curricular group added another level for analysis. A different kind of curricular intervention is added. In addition to classes that have woven debates into the class, classes that are fundamentally about argumentation and debate were also assessed. This should mark a different but more in-depth curricular intervention. For rating the student responses, the same rater and rating process utilized in Study 2 was followed here for Study 3 and Study 4.

Participants

All of the samples administered the argumentation education instrument represent convenient samples of college students recruited for participation.

Study 1

The participants (n=6) are scholars and intercollegiate debate coaches across the United States considered subject matter experts in argumentation. The author generated a list of 19 possible participants to send the survey to for reviewing the argumentation

education assessment instrument. The list was generated based on two things. First, the author generated the list based on his perception that these individuals are among the leading scholars and practitioners of argumentation and debate, having himself been a member of this discipline for over ten years. And second, the list included participants whom the author considered were likely to respond.

Study 2

Three convenient samples of college students were recruited for participation. First, a control group (n=46) was identified of students enrolled in entry-level communication courses at a major Mid Atlantic university. These students are required to participate in a research pool as a grade for their course. Students may have opted into this particular research pool option for any number of reasons. Second, a college curriculum intervention group (n=41) was identified from students enrolled in two different courses at the same Mid Atlantic university where the instructional faculty intentionally integrated argument education into the classroom. Before the beginning of the semester, the faculty members were consulted, through workshops and individually, on how to implement argument education for their course. One course was an entry-level communication course and the other a health sciences class. While students from both classes were recruited for participation, all but one student in the curricular intervention group was from the communication course. Faculty worked on curriculum adapted to their discipline and course restraints. While different and specific for each course, the curricular intervention for argumentation was consistent in that certain aspects of argument education were present throughout all of them. For example, each curricular intervention involved group collaboration, public speaking, argumentation, research,

decision-making, and perspective taking. Third, a debate extra curricular intervention group (n=6) was identified from students who actively compete at college policy debate tournaments on the National Debate Tournament (NDT)/Cross Examination Debate Association (CEDA) circuit. College policy debate coaches were recruited to have their program participate in this study based on willingness to encourage student-debater participation and likelihood to follow-through. College policy debate is the format of debate selected because this format emphasizes the skillsets targeted here by argument education, for example group collaboration, argumentation, and research.

The students in the control group and curriculum group attend a mid-sized master's level mid-Atlantic institution of higher education. The competitive policy debate group students attend a variety of institutions of higher education, from private to public and community college through Ivy League. The students self-reported demographic information such as classification in school, major, race, and gender identity. The samples were largely white, female, and not international students (Tables 2, 3, and 4). The control group was slightly more diverse racially and had more male students. Both the control and curricular intervention groups had little to no debate experience while the extra-curricular debate students all had prior debate experience (Table 5). Finally, both the control and curricular intervention groups were made up mostly of students in their first year of college (Table 6).

Table 2
Study 2 sample by race

Race	Control (n=46)	Curricular (n=41)	Debate (n=6)	Total (n=93)
American Indian	2	0	1	3
Asian	6	4	0	10

Black	4	2	0	6
Hispanic	4	1	1	6
Native Hawaiian	0	1	0	1
White	32	32	5	69
Another	0	0	0	0
Prefer not to answer	1	1	0	2

* Totals may be higher because participants can check multiple answers

Table 3
Study 2 sample by gender identity

Gender identity	Control (n=46)	Curricular (n=41)	Debate (n=6)	Total (n=93)
Female	22	33	3	58
Male	25	6	3	34
Transgender	0	0	0	0
Queer	0	1	0	1
Another	0	0	0	0
Prefer not to answer	0	1	0	0

* Totals may be higher because participants can check multiple answers

Table 4
Study 2 sample by international student status

International student status	Control (n=46)	Curricular (n=41)	Debate (n=6)	Total (n=93)
Yes	4	0	0	4
No	41	40	6	87
Prefer not to answer	1	1	0	2

Table 5
Study 2 sample by prior debate experience

Prior debate experience	Control (n=46)	Curricular (n=41)	Debate (n=6)	Total (n=93)
No experience	30	15	0	45
High school class debates	14	24	2	40
High school competitive debate	2	1	3	6
College class debates	3	12	0	15
College competitive debate	3	1	6	10
Other	0	2	0	2
Prefer not to answer	2	1	0	3

* Totals may be higher because participants can check multiple answers

Table 6
Study 2 sample by college standing

College standing	Control (n=46)	Curricular (n=41)	Debate (n=6)	Total (n=93)
First year student	42	38	1	81
Sophomore	0	1	1	2
Junior	0	0	1	1
Senior	2	1	1	4
Graduate student	0	0	2	2
Prefer not to answer	2	1	0	3

Study 3

The control group (n=182) is the same type of sample represented in Study 1 as the control group. These are students from a Mid-Atlantic university enrolled in an entry-level communication course who are required to participate in a research pool for their course grade. A different type of curricular intervention was added and treated separately for study 3. Curricular intervention 1 (n=157) is similar to the curricular intervention participants in Study 1. These are students who are enrolled in two sections of a class, where their faculty member has integrated debate and argument education into the classroom. This faculty member has worked with the author to design and implement argumentation based debate activities into their class. Again, these participants were from the same Mid-Atlantic university. The second curricular group is constituted by participants from a different form of intervention. Curricular intervention 2 participants (n=72) are students enrolled in an argumentation and debate class. The faculty teaching these classes were recruited via social media. Furthermore, emailed the study's author the syllabus and other information about how they integrate argumentation into their course experience. The debate extra-curricular intervention group (n=36) were students who

compete in a collegiate debate format. As in Study 2, these participants were recruited from the NDT/CEDA college policy debate circuit.

Like study 2, the students from the control and curricular 1 intervention group attend a mid-sized master's level mid-Atlantic institution. The participants from curricular 2 and debate attend varying institutions of higher education from across the U.S. The student participants across all groups self-reported all of their demographic information. The control and curricular 1 groups were largely white, female, and not international students (Table 7, Table 8, and Table 9). Note that the pre/post groups were collapsed for the reporting of their demographic information. Students from the curricular 2 and debate groups were more diverse in their reported race and gender identity. The control, curricular 1, and curricular 2 groups reported little to no prior debate experience (Table 10) while the debate group did not complete that part of the survey. Finally, participants in the control group were mostly first year students while participants in the curricular groups were generally sophomores and juniors (Table 11). The debate group had students from across academic standings.

Table 7
Study 3 sample by race

(n=447) Race	Control (n=182)	Curricular 1 (n=157)	Curricular 2 (n=72)	Debate (n=36)	Total
American Indian	2	0	1	0	3
Asian	12	2	13	0	27
Black	10	4	11	2	27
Hispanic	8	7	24	14	53
Native Hawaiian	0	1	1	1	3
White	156	150	28	15	349
Another	0	0	3	4	7
Prefer not to answer	1	0	3	0	4

* Totals may be higher because participants can check multiple answers

Table 8
Study 3 sample by gender identity

	Control (n=182)	Curricular 1 (n=157)	Curricular 2 (n=72)	Debate (n=36)	Total (n=447
Gender identity					
Female	131	151	49	17	348
Male	50	5	17	19	91
Transgender	0	0	3	0	3
Queer	1	0	0	0	1
Another	0	0	1	0	1
Prefer not to answer	1	1	2	1	5

* Totals may be higher because participants can check multiple answers

Table 9
Study 3 sample by international student status

	Control (n=182)	Curricular 1 (n=157)	Curricular 2 (n=72)	Debate (n=36)	Total (n=447)
International student status					
Yes	3	0	1	2	6
No	177	156	69	34	436
Prefer not to answer	2	0	2	0	4

Table 10
Study 3 sample by prior debate experience

	Control (n=182)	Curricular 1 (n=157)	Curricular 2 (n=72)	Debate (n=36)	Total (n=447)
Prior debate experience					
No experience	111	70	48	-	229
High school class debates	66	68	15	-	149
High s. competitive debate	5	4	0	-	9
College class debates	3	32	10	-	45
College competitive debate	2	0	1	-	3
Other	3	0	1	-	4
Prefer not to answer	3	1	1	-	5

* Totals may be higher because participants can check multiple answers

Table 11
Study 3 sample by college standing

	Control (n=182)	Curricular 1 (n=157)	Curricular 2 (n=72)	Debate (n=36)	Total (n=447)
College standing					
First year student	172	0	3	6	181
Sophomore	7	97	47	6	157

Junior	1	51	12	9	73
Senior	1	8	9	15	33
Graduate student	0	0	0	0	0
Prefer not to answer	1	0	1	0	2

Study 4

Participants in study 4 were very similar in number and demographic make-up to the participants from study 3 (Tables 12-16). Debate was the exception group as it was significantly smaller for study 4 (n=14). The control group and curricular 1 groups were largely female, white, and not international students. The curricular 2 and debate groups had a little more diverse representation, especially for race. Most participants across the control and curricular groups had little to no experience with debate prior to the administration of the survey. The academic class standing, again, mirrored study 3 with most of the control being first year students, while the other groups were composed of largely sophomores and juniors.

Table 12
Study 4 sample by race

	Control (n=170)	Curricular 1 (n=137)	Curricular 2 (n=88)	Debate (n=14)	Total (n=409)
Race					
American Indian	0	0	3	2	5
Asian	12	4	6	2	24
Black	8	0	21	4	33
Hispanic	9	5	10	11	35
Native Hawaiian	2	0	0	0	2
White	150	129	49	2	330
Another	2	2	2	0	6
Prefer not to answer	1	0	3	3	7

* Totals may be higher because participants can check multiple answers

Table 13
Study 4 sample by gender identity

	Control (n=170)	Curricular 1 (n=137)	Curricular 2 (n=88)	Debate (n=14)	Total (n=409)
Gender identity					

Female	116	131	64	8	319
Male	53	4	21	2	80
Transgender	0	2	0	0	2
Queer	0	0	1	2	3
Another	1	0	1	0	2
Prefer not to answer	0	0	2	2	4

* Totals may be higher because participants can check multiple answers

Table 14
Study 4 sample by international student status

	Control (n=170)	Curricular 1 (n=137)	Curricular 2 (n=88)	Debate (n=14)	Total (n=409)
International student status					
Yes	4	0	1	2	7
No	165	135	85	12	397
Prefer not to answer	1	2	2	0	5

Table 15
Study 4 sample by prior debate experience

	Control (n=170)	Curricular 1 (n=137)	Curricular 2 (n=88)	Debate (n=14)	Total (n=409)
Prior debate experience					
No experience	102	63	74	-	239
High school class debates	59	60	1	-	120
High s. competitive debate	9	2	3	-	14
College class debates	13	25	7	-	45
College competitive debate	0	1	1	-	2
Other	2	0	2	-	4
Prefer not to answer	2	0	2	-	44

* Totals may be higher because participants can check multiple answers

Table 16
Study 4 sample by college standing

	Control (n=170)	Curricular 1 (n=137)	Curricular 2 (n=88)	Debate (n=14)	Total (n=409)
College standing					
First year student	156	4	0	4	164
Sophomore	8	104	18	3	133
Junior	6	25	40	2	73
Senior	0	4	28	5	37
Graduate student	0	0	0	0	0
Prefer not to answer	0	0	2	0	2

Analyses

Four types of analyses were used throughout the research studies. First, g-theory was used to analyze, determine, and assess a measurement estimate of inter-rater reliability for the argumentation education assessment instrument. G-studies were run to determine the g-coefficient for each instrument use in Study 1 and Study 3. Additionally, d-studies were conducted to identify the different g-coefficient possibilities should different facet levels be used. These analyses helped answer Hypothesis 1.

This study used a three-facet (or possible sources of error) design to conduct the g- and d-studies. Persons, raters, and items were each considered a facet. The facets were all treated as random because the universe of generalization is all possible students (or potential raters) in U.S. higher education. This allows the maximum flexibility and use of the instrument. A fixed facet would have limited the generalizability because the instrument could have only been used in certain conditions, like a set group of students and specific raters. Furthermore, the design had raters and items fully crossed within persons. This meant that all raters rated all items for all persons.

Hypothesis 1. The argumentation education assessment instrument will yield a generalizability-coefficient greater than 0.70.

Second, descriptive statistics were utilized to analyze the information collected from the argumentation subject matter experts in Study 2's instrument review survey. Third, one-way analysis of variance (ANOVA) was used to compare the item scores on the argumentation education assessment instrument between the control, curricular intervention, and extra curricular debate intervention groups. Furthermore, eta squared

was reported for an effect size for the ANOVA test and Cohen's *d* for any difference between group means. This set of analyses will help answer Hypothesis 2a and 2b.

Hypothesis 2a. Students participating in a collegiate curricular intervention where debate pedagogy has been integrated into the curriculum will score higher on an argumentation education instrument than students in a control group.

Hypothesis 2b. Students participating in a collegiate extra-curricular debate intervention will score higher on an argumentation education instrument than students in a control group and students in the collegiate curricular intervention.

Fourth, the pre/post within group means was analyzed using a one-way analysis of variance (ANOVA) to identify statistically significant differences for interactions between group, pre/post test, and the five argument scale items. Eta-squared is reported for the ANOVA tests and Cohen's D reported as an effect size for differences between paired sample pre/post group means on the argumentation education assessment instrument. These analyses will help answer Hypothesis 3a and 3b.

Hypothesis 3a. Students participating in a collegiate curricular intervention where debate pedagogy has been integrated into the curriculum will score higher on an argumentation education instrument than before their intervention.

Hypothesis 3b. Students participating in a collegiate extra-curricular debate intervention will score higher on an argumentation education instrument than before their intervention.

CHAPTER 4

Results

The results in this chapter align three of the four studies conducted with the four hypotheses. Recall that Study 1 asked subject matter experts in the field of argumentation and debate to review the argumentation assessment instrument. This data was used as part of the case for instrument validation rather than to answer any of the hypotheses. In Study 2, the instrument was piloted at the end of the spring 2016 semester as a post-test for a control group, argumentation curricular intervention group, and debate extra-curricular group. Study 3 continued to use the argumentation assessment instrument, but expanded on study 2 by increasing the sample size of each group and adding a pre-test in addition to a post-test, both taking place during the fall 2016 semester. Finally, study 4 was a replication study that took place during the spring 2017 semester. Again, three samples were used (control, argumentation curricular intervention, debate extra-curricular) for a pre and post-test, but the argumentation assessment instrument added an additional item to pilot measuring affective argumentation identification. The instrument's reliability data from Study 2, 3, and 4 addresses Hypothesis 1. The participant scores between samples on the instrument from Study 2, Study 3, and Study 4 addresses Hypotheses 2a and 2b. Hypotheses 3a and 3b are answered utilizing the participant scores within samples on the instrument from Study 3 and Study 4.

Study 1: Argument subject matter expert review. This study gathered subject matter expert review evidence toward validating the argumentation education assessment instrument within the higher education context. First, a survey was sent to argumentation subject matter experts to review the definition, rubric and prompts that make the

instrument. The results of the subject matter expert review skewed toward agreeing or strongly agreeing that the definition, rubric, and scenario prompts were acceptable and aligned with one another (Table 17). Five out of six experts agreed that David Zarefsy's definition of argumentation is an acceptable foundational definition of argumentation. While there was general agreement that this definition was acceptable, two experts did respond that the definition ignored "the influence of audience" and left "unexamined the question of reasonableness."

Table 17
Study 1 argumentation instrument subject matter expert review

Items	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
Acceptable definition of argumentation?					
	0	0	1	3	2
Foundational argumentation skill?					
Bias	0	0	0	2	3
Prioritization	0	0	0	1	4
Construction	0	0	0	4	1
Evaluation	0	0	0	2	3
Utilization	0	0	1	2	2
Rubric performance criteria clear and reflective?					
Bias	0	0	1	2	2
Prioritization	0	0	1	1	3
Construction	0	0	3	1	1
Evaluation	0	0	1	2	2
Utilization	0	0	2	1	2
Scenario prompt aligns with rubric criteria?					
Bias	0	0	0	0	4
Prioritization	0	0	0	2	2
Construction	0	1	0	2	1
Evaluation	0	0	0	2	2
Utilization	0	0	0	1	3

* Total N varies as participants dropped out of survey

The subject matter experts also generally agreed that the five behaviors identified on the rubric are foundational argumentation skills. Four of the five skills received five

out of five experts agreeing or strongly agreeing that the skill was foundational to argumentation. Utilizing argument in a situation received the weakest support, only generating four agreeing or strongly agreeing and one neither agree or disagree. Three of the argumentation experts said they would not remove of any of the five skills from the rubric while one identified logical argument construction and one identified argument utilization in a situation. None of the experts listed other performance criteria that should be added as a foundational argumentation skill.

The experts were also asked about whether the performance criteria rubric is clear and reflective of the solicited performance. Three of the rubric criteria received four out of five agreement or strong agreement from the experts. Argument construction received two expert agreements while argumentation utilization received three. None of the five performance criteria received any disagreement about being clear and reflective of the performance criteria. Three of the argumentation experts provided qualitative feedback about the performance criteria on the rubric. The feedback ranged from questions about what reasonable means in the context of explicit warrants to how much logical argument construction is influenced by the work of Stephen Toulmin.

Finally, the subject matter experts were asked if the scenario prompts aligned with each of their respective rubric categories and performance criteria. Four out of the five scenario prompts received consensus agreement from the four experts still responding, either agreeing or strongly agreeing. Only the logical construction of argument received a disagree from one expert. Three experts provided qualitative feedback to the prompts, one saying “these are very good” while the other two asking for more detail on the logical construction of argument scenario. All experts had the opportunity for general feedback

after thanking them at the end of the instrument review. Only one expert responded, saying "Great work – eager to see your work when this is complete."

Hypothesis 1: Argumentation education assessment instrument generalizability-coefficient. Generalizability and decision studies were calculated for Study 2, Study 3, and Study 4 to answer Hypothesis 1. While H1 relies on a g-coefficient to provide evidence for the instrument's scoring reliability, other inter-rater reliability coefficients were also calculated to provide context for interpreting the g-coefficient and assessing the reliability of the scores across all three studies. The g-coefficient was used for a measurement estimate, while Pearson's Correlation represents a consistency estimate. Percent agreement, Cohen's Kappa, Pearson's Correlation, Gwet's AC1, Scott's Pi, Krippendorff's Alpha, and Brennan-Prediger reflect measures of consensus estimate. A major difference among these consensus estimates centers on how each defines and calculates agreement and/or chance. For example, Scott's Pi theorizes chance by assuming that a rater at random could potentially assign a score in any given cell while Gwet's AC1 articulates chance as a function of how hard versus easy subjects are to rate.

Study 2, spring 2016. Generalizability theory was used to calculate a g-coefficient for a measurement estimate, both for the individual item (5 items) and the total score (sum of score on the five items). A g-study was run across all groups (n=93) and items (n=5), utilizing a P/RI design (Table 18). The g-coefficient for this g-study was 0.43. Follow-up g-studies were run analyzing each of the two larger groups (control and curricular intervention, respectively n=46 and n=41) because the original coefficient for all groups seemed low. The person by item variance was of particular interest because it represented the majority of the variance in the original g-study.

Table 18
G-study for all groups (P/RI)

Source of Variance	SS	DF	MS	Absolute Error Variance	Percent of Error Variance
Person	138.93	92	1.51	…	
Rater	0.02	1	0.02	0	0%
Item	75.18	4	18.79	0.02	17.5%
Person X Rater	7.38	92	0.08	0	0%
Person X Item	319.82	368	0.87	0.08	73.8%
Rater X Item	3.04	4	0.76	0	0.7%
Person X Rater X Item	31.56	368	0.09	0.01	8.1%
Total	579.93	929		0.11	100%
G-Coefficient 0.43					

The follow-up g-study for the control group (Table 19) and curricular intervention group (Table 20) revealed similar results. Both g-studies used the same design, P/RI. The control group g-study had a coefficient of 0.45, with 70.3% of the variance due to person by item interaction. The g-study for the curricular intervention group had a g-coefficient of 0.44 with 73.5% of variance resulting from the person by item interaction.

Table 19
G-study for control group (P/RI)

Source of Variance	SS	DF	MS	Absolute Error Variance	Percent of Error Variance
Person	67.92	45	1.51	…	
Rater	0	1	0	0	0%
Item	35.55	4	8.89	0.02	16.8%
Person X Rater	4.5	45	0.10	0	0
Person X Item	150.85	180	0.84	0.07	70.3%
Rater X Item	1.42	4	0.36	0	0.5%
Person X Rater X Item	22.58	180	0.13	.01	12.4%
Total	282.82	459		0.10	100%
G-Coefficient 0.45					

Table 20
G-study for curricular intervention group (P/RI)

Source of Variance	SS	DF	MS	Absolute Error Variance	Percent of Error Variance
Person	63.3	40	1.58	…	
Rater	0.02	1	0.02	0	0%
Item	39.67	4	9.92	0.02	19%
Person X Rater	2.88	40	0.07	0	1.6%
Person X Item	138.53	160	0.87	.08	73.5%
Rater X Item	1.99	4	0.50	0	1%
Person X Rater X Item	8.61	160	0.05	0.01	4.9%
Total	255	409		0.11	100%

G-Coefficient 0.44

Given the high amount of variance attributed to the items throughout each g-study, one final g-study was run for the total score without the five items differentiated (Table 21). This g-study only analyzed the person and raters, using a P/R design. The g-coefficient for this study looking only at the raters and overall score was 0.95. The person by rater interaction accounted for 100% of the variance within this g-study.

Table 21
G-study for all groups (P/R)

Source of Variance	SS	DF	MS	Absolute Error Variance	Percent of Error Variance
Person	723.87	93	7.78	…	
Rater	0.19	1	0.19	0	0%
Person X Rater	37.81	3	0.41	0.20	100%
Total	761.87	187			100%

G-Coefficient 0.95

Finally, a d-study was run taking advantage of the ability for generalizability theory to project g-coefficients into the universe with different facet elements (Table 22). The original g-study had a g-coefficient of 0.43 with five items and two raters. The d-study varied both the rater and item facets. The items varied from five to seven items and the raters from one rater to five raters. The g-coefficients across all possible facet

combinations ranged from 0.40 to 0.53. Using one rater and five items resulted in the lowest g-coefficient, 0.40, while utilizing five raters and seven items increased the g-coefficient 0.10 to 0.53 over the current study of two raters and five items.

Table 22
D-study for varying raters and items (P/RI)

	1 rater	2 raters	3 raters	4 raters	5 raters
5 items	0.40	0.43	0.43	0.43	0.44
6 items	0.45	0.47	0.48	0.48	0.49
7 items	0.49	0.51	0.52	0.52	0.53

Multiple inter-rater reliability coefficients were calculated to reflect the other ways inter-rater reliability is theorized and provide additional context to the g-coefficients here (Table 23). Both item and total score on the argumentation assessment instrument were used throughout to better explore and understand the instrument's scoring reliability. In addition to the total score, a weighted total was calculated. Weights were added on the total but not the items because the range of the total (0-10) varied more than on each item (0-2). Furthermore, exact agreement was of more concern for each item because of the meaningful difference between a 0 and a 2. While exact agreement on the total was of less concern because the differences were less meaningful, for example between a 9 and an 8.5.

Table 23
Argumentation instrument inter-rater reliability coefficients and rater agreement

Coefficient	Arg Eval	Arg Util	Arg Bias	Arg Const	Arg Prior	Total*	Weighted Total
Pearson Correlation	0.75	0.81	0.79	0.74	0.99	0.90	…
Cohen's Kappa	0.67	0.64	0.66	0.61	0.98	0.45	0.75
Gwet's AC1	0.79	0.74	0.66	0.71	0.99	0.50	0.86
Scott's Pi	0.67	0.63	0.66	0.60	0.98	0.45	0.75
Krippendorff's Alpha	0.67	0.63	0.66	0.60	0.98	0.46	0.75
Brennan-Prediger	0.76	0.71	0.66	0.78	0.98	0.49	0.84
Percent Agreement	0.84	0.81	0.77	0.78	0.99	0.54	0.94

* Total is the sum of all of the items (0-10 scale rather than 0-2 scale for items)

Percent agreement was calculated to reflect a consensus estimate. The items ranged from 77% to 99% percent agreement between the two raters and the total was 54% percent agreement. The weighted total for percent agreement was 94%. Cohen's Kappa, another measure of consensus estimate, ranged from .67 to .98 for the items alone. Kappa was 0.75 for the weighted total while the unweighted total had a Kappa of 0.50.

Cronbach's alpha was also calculated because of how the different inter-rater reliability coefficients were emerging. On the one hand, the g-studies were showing a high percentage of variance due to the item and item interactions, resulting in a low g-coefficient. On the other hand, the other benchmarks for inter-rater reliability were generally above the 0.70 recommended threshold for acceptable reliability, demonstrating fairly reliable scores from the raters on the instrument. Cronbach's alpha for the five items on the instrument was .425, with inter-item correlations never above 0.258. This low Cronbach's alpha identifies a trend with the inter-rater reliability data that high amounts of variance resides within item scores on the instrument rather than the raters scoring.

Study 3, fall 2016. A g-coefficient was calculated for the fall 2016 overall sample (n=447) across all items (n=5) running a g-study that utilized a P/IR design (Table 24). The g-coefficient for this study was 0.38 with most of the variance clustering around the item facet. Item variance alone represented 16.5% of variance while the person by item interaction accounted for 72% of total variance. No additional g-studies were conducted

to test and try to explain the high item variance because the results closely mirrored the initial spring 2016 pilot study.

Table 24
G-study for all groups (P/IR)

Source of Variance	SS	DF	MS	Absolute Error Variance	Percent of Error Variance
Person	447.51	446	1.00	…	
Item	225.60	4	56.40	0.01	16.5%
Rater	2.10	1	2.10	0.00	0.40%
Person X Item	1096.00	1784	0.61	0.05	72.0%
Person X Rater	36.40	446	0.08	0.00	0.70%
Item X Rater	3.44	4	0.86	0.00	0.20%
Person X Item X Rater	136.56	1784	0.08	0.01	10.2%
Total	1947.61	4469		0.07	100%
G-Coefficient	0.38				

A decision study was run to determine how altering raters and items might impact the g-coefficient (Table 25). The d-study found that increasing the number of items had the most effect on the g-coefficient, ranging from an increased coefficient of 0.43-0.44. Increasing raters had minimal impact on the g-coefficient, only increasing from 0.38 to 0.40 by doubling the number of raters (two to four).

Table 25
D-study for varying raters and items (P/RI)

	2 rater	3 raters	4 raters
5 items	0.38	0.39	0.40
6 items	0.43	0.44	0.44

Other inter-rater reliability coefficients were calculated to include the consensus and consistency estimates (Table 26). For this study, only the items were analyzed at the item level because the total score reflected a sum of the items rather than a value with new insight. The argument construction item recorded the lowest reliability coefficients while argument prioritization reported the highest. All of the coefficients, except for

argument construction, were right around or well above the 0.70 threshold. Percent agreement was the measure that consistently had the highest coefficient while Cohen's Kappa was generally the lowest of the measures. Cronbach's alpha was calculated at 0.386 across the five items on the instrument, with 0.187 as the highest correlation on the inter-item correlation matrix.

Table 26
Argumentation instrument inter-rater reliability coefficients and rater agreement

Coefficient	Arg Eval	Arg Util	Arg Bias	Arg Const	Arg Prior
Pearson Correlation	0.89	0.76	0.92	0.72	0.94
Cohen's Kappa	0.80	0.68	0.84	0.48	0.88
Gwet's AC1	0.91	0.74	0.84	0.61	0.91
Scott's Pi	0.80	0.68	0.84	0.47	0.88
Krippendorff's Alpha	0.80	0.68	0.84	0.47	0.88
Brennan-Prediger	0.89	0.72	0.84	0.57	0.90
Percent Agreement	0.93	0.82	0.89	0.71	0.94

Study 4, spring 2017. A generalizability coefficient was calculated for the spring 2017 sample (n=409). A g-study was run, using the P/IR design, that resulted in a .027 g-coefficient (Table 27). Moreover, the item facet was the source for most of the variance like in study 2 and study 3. The item facet alone represented 6.5% of the overall variance while the person by item interaction reflected the largest source of overall variance, 84.7% respectively. A decision-study was run to calculate the impact of varying items and raters on the g-coefficient (Table 28). Again, varying the items had the greatest impact of the coefficient. Increasing the raters from 2 to 4 only increase the g-coefficient .01, while increasing the items from 5 to 7 increased the coefficient .08 to 0.35.

Table 27
G-study for all groups (P/IR)

Source of Variance	SS	DF	MS	Absolute Error Variance	Percent of Error Variance
Person	493.41	408	1.21	…	
Item	109.53	4	27.38	0.01	6.5%
Rater	2.85	1	2.85	0.00	0.4%
Person X Item	1436.87	1632	0.88	0.08	84.7%
Person X Rater	30.35	408	0.07	0.00	0%
Item X Rater	5.44	4	1.36	0.00	0.3%
Person X Item X Rater	125.36	1632	0.08	0.01	8.1%
Total	2203.81	4089		0.09	100%
G-Coefficient	0.27				

Table 28
D-study for varying raters and items (P/RI)

	2 rater	3 raters	4 raters
5 items	0.27	0.28	0.28
6 items	0.31	0.32	
7 items	0.35		

Multiple inter-rater reliability coefficients were calculated to supplement the g-coefficient with consistency and consensus measures (Table 29). The item argument construction received the lowest reliability coefficients while the argument prioritization item received the highest reliability scores. Argument construction was the item that received the most coefficients below 0.70, while argument evaluation had some hovering around 0.70. Cohen's Kappa, again, represent the lowest measure of inter-rater reliability while percent agreement reported the highest measure. The Cronbach's alpha for the five argument items on this instrument was 0.272 and 0.153 was the highest correlation between items on the inter-item correlation matrix.

Table 29
Argumentation instrument inter-rater reliability coefficients and rater agreement

Coefficient	Arg Eval	Arg Util	Arg Bias	Arg Const	Arg Prior
Pearson Correlation	0.72	0.82	0.89	0.61	0.98
Cohen's Kappa	0.68	0.78	0.80	0.47	0.95
Gwet's AC1	0.85	0.82	0.81	0.57	0.96
Scott's Pi	0.68	0.79	0.80	0.46	0.95
Krippendorff's Alpha	0.68	0.79	0.80	0.46	0.95
Brennan-Prediger	0.82	0.81	0.81	0.53	0.96
Percent Agreement	0.88	0.88	0.87	0.69	0.98

The reliability coefficients across all three studies demonstrate a few important consistent trends. First, the g-studies and d-studies point toward low g-coefficients. Even significantly increasing the number of raters and items does not bring the g-coefficient close to the 0.70 benchmark for an acceptable reliability coefficient. Furthermore, all three studies found item and item interactions to be the highest source of variance, not the raters. Second, the other inter-rater reliability coefficients were fairly consistent in hovering acceptably around or well above 0.70. Moreover, the differences in the coefficient across argument items were also consistent across the three studies, with argument construction being among the lowest. Third, Cronbach's alpha was low across all three studies, with very low correlations in each respective inter-item correlation matrix. The data collected does not support Hypothesis 1 that the argument education assessment instrument will yield a g-coefficient above 0.70. The implications for these trends are discussed in the next chapter.

Hypothesis 2a and 2b: Comparing between group (varying levels of argument education curricular integration) scores on the argumentation education assessment instrument. Both versions of hypothesis 2 were concerned with assessing the differences

between groups on the argumentation education assessment instrument. For version 2a, it was hypothesized that students in a college class where some form of argument education had been integrated into the course curriculum would report higher scores than students in a college class identified as a control group. For version 2b, it was hypothesized that students actively participating in an extra-curricular debate organization would report higher scores on the argumentation education assessment instrument than either students from the control group or the collegiate curricular intervention group. To address these hypotheses, a one-way analysis of variance (ANOVA) was run for each different study, with the item scores and total score as the dependent variables and the group involvement as the independent variable. The Kruskal-Wallis test was also run for each study to check the results of the ANOVA because of possible concerns with sample distribution.

Study 2, spring 2016. The scores for each group by item and total score are reported in Table 30. A one-way analysis of variance (ANOVA) was conducted to compare the differentiated item and total score in the control (n=46), curricular intervention (n=41), and extra-curricular intervention groups (n=6). A major assumption for the ANOVA test is that there is homogeneity of variance (Field, 2013, p. 442). Levene's test was run to determine if there were statistically significant differences of variance within groups for each argument instrument item and the total instrument score. All tests for homogeneity of variance were non-significant, with the closest one being argument evaluation having a p-value of .077. For the ANOVA proper, there was a statistically significant effect for the argument evaluation item across the three groups *($F(2,90)=3.25$, $p.=.04$, $\eta^2=.07$).* Post hoc comparisons using the Tukey HSD test indicated the mean score for the debate extra-curricular group (M=1.67, SD=0.52) was

significantly different than both the curricular intervention (M=1.10, SD=0.45) and the control group (M=1.14, 0.56). A Kruskal-Wallis H test showed that there were not any statistically significant differences in argumentation education scores between the different argument curriculum interventions. Although the test showed that the closest statistically significant different was on argument evaluation, $x^2(2) = 5.646$, $p = .059$, with a mean rank of 46.57 for the control group, 44.24 for the curricular group, and 69.17 for the debate extra-curricular group.

Table 30
Scores on the argumentation instrument

Item (scored 0, 1, 2)	Control (n=46) (M, SD)	Curricular (n=41) (M, SD)	Extra-Curricular (n=6) (M, SD)	Total (n=93) (M,SD)
Argument Evaluation	1.14, .56*	1.10, 0.45**	1.67, 0.52* **	1.16, 0.53
Argument Utilization	0.49, 0.70	0.46, 0.65	0.83, 0.98	0.50, 0.69
Argument Bias	0.86, 0.71	0.85, 0.84	1.50, 0.55	0.90, 0.77
Arg Construction	1.17, 0.56	1.27, 0.54	1.50, 0.55	1.24, 0.55
Arg Prioritization	1.25, 0.91	1.30, 0.95	1.00, 0.89	6.50, 1.05
Total	4.91, 1.95	4.99, 1.99	6.5, 1.05	5.04, 1.95

* Significant between-subject ANOVA tests
** Significant between subject ANOVA tests

None of the other individual items or total instrument score had statistically significant effects. But while there were no statistically significant effects between the three groups for the other items and total score, practical significance was also calculated (see Table 31). The effect sizes were calculated because sample size may limit ability to detect statistically significant differences. And within this study, the sample size of the extra-curricular intervention (n=6) may have impacted the ability to show significant differences while the means and standard deviations suggested that possibly meaningful differences existed between groups. The extra-curricular group did have at least a

moderate effect size when compared individually with both the control and curricular intervention in four of the five items and the total score.

Table 31
Group comparison effect sizes on the argumentation instrument

Groups compared by item	Eta-squared	P-value	Cohen's D	Magnitude
Argument evaluation	0.07	0.04	-	Medium
Control vs Extra-Curricular	-	-	0.97	Large
Curricular vs Extra Curricular	-	-	1.17	Large
Control vs Curricular	-	-	0.09	None
Argument Utilization	0.02	0.47	-	Small
Control vs Extra-Curricular	-	-	0.40	Moderate
Curricular vs Extra Curricular	-	-	0.44	Moderate
Control vs Curricular	-	-	0.04	None
Argument Bias	0.04	0.14	-	Small
Control vs Extra-Curricular	-	-	1.00	Large
Curricular vs Extra Curricular	-	-	0.91	Large
Control vs Curricular	-	-	0.01	None
Argument Construction	0.02	0.35	-	Small
Control vs Extra-Curricular	-	-	0.59	Moderate
Curricular vs Extra Curricular	-	-	0.43	Moderate
Control vs Curricular	-	-	0.17	Small
Argument Prioritization	0.01	0.75	-	Small
Control vs Extra-Curricular	-	-	0.28	Small
Curricular vs Extra Curricular	-	-	0.33	Small
Control vs Curricular	-	-	0.06	None
Argument Total	0.04	1.66	-	Small
Control vs Extra-Curricular	-	-	1.01	Large
Curricular vs Extra Curricular	-	-	0.95	Large
Control vs Curricular	-	-	0.04	None

Study 3, fall 2016. In study 3, recall that the sample numbers were increased, an additional argument curriculum group added, and a pre-test was administered for each group in addition to the post-test. The scores for each group by item and overall total score are reported in Table 32. An ANOVA was conducted to compare the group scores

(Control, Curricular 1, Curricular 2, and Debate) by item and overall total score. Again, Levene's test was run to test for homogeneity of variances and found that the test was statistically significant for argument

Table 32
Scores on the argumentation instrument

Group	Evaluation	Utilization	Bias	Construction	Prioritization	Total
Control Pre (n=88) (M, SD)	1.17, 0.41	0.89, 0.55	1.06, 0.78	1.50, 0.41*	1.52, 0.56	6.13, 1.50
Control Post (n=94) (M, SD)	1.07, 0.33	0.94. 0.48	0.94, 0.77	1.40, 0.56	1.56, 0.52	5.92, 1.49
Curricular_1 Pre (n=69) (M, SD)	1.13, 0.42	0.82, 0.56	1.12, 0.76	1.46, 0.50	1.43, 0.65	5.96, 1.67
Curricular_1 Post (n=88) (M, SD)	1.17, 0.42	0.67, 0.66	1.16, 0.74	1.25, 0.53*	1.47, 0.60	5.72, 1.63
Curricular_2 Pre (n=39) (M, SD)	1.08, 0.51	0.87, 0.50	1.18, 0.75	1.51, 0.47	1.45, 0.62	6.09, 1.70
Curricular_2 Post (n=33) (M, SD)	1.29, 0.59	0.94, 0.75	1.06, 0.79	1.46, 0.52	1.18, 0.66	5.92, 1.78
Debate Pre (n=19) (M, SD)	1.32, 0.48	1.03, 0.63	1.08, 0.82	1.42, 0.67	1.34, 0.67	6.18, 1.74
Debate Post (n=17) (M, SD)	1.29, 0.56	1.00, 0.71	1.06, 0.77	1.56, 0.46	1.38, 0.55	6.29, 1.59
Total Pre (n=215) (M, SD)	1.15, 0.43	0.87, 0.55	1.10, 0.77	1.48, 0.47	1.46, 0.61	6.07, 1.57
Total Post (n=232) (M, SD)	1.16, 0.43	0.84, 0.62	1.05, 0.76	1.36, 0.54	1.46, 0.58	5.87, 1.59
Total (n=447) (M, SD)	1.16, 0.43	0.86, 0.59	1.07, 0.76	1.42, 0.51	1.46, 0.59	5.97, 1.58

* Significant between-subject ANOVA tests at the .05 level

evaluation, argument utilization, and argument construction. This suggests that the variances were statistically different from one another and a statistical correction was needed to overcome the violated ANOVA assumption. Welch's F was used as the corrected F-ratio because "The Welch test seems to fare the best except when there is an extreme mean that has a large variance" (Field 2013, p. 443). Utilizing Welch's F, there was a statistically significant effect for the argument construction item across the eight groups *(F(7,439)=2.193, p.=.04, η^2=.033)*. Post hoc comparisons using the Tukey HSD test indicated the mean score for the control group pre-test (M=1.50, SD=0.41) was significantly different than the curricular 1 intervention post-test (M=1.25, SD=0.53). A Kruskal-Wallis H test showed that there were was a statistically significant difference in scores on the argument construction item, $x^2(2) = 14.239$, $p = .047$; with a mean rank of 239.15 for the control group pre-test, 224.64 for the control group post-test, 231.70 for the curricular 1 intervention group pre-test, 182.05 for the curricular 1 intervention group post-test, 243.12 for the curricular 2 intervention group pre-test, 231.30 for the curricular 2 intervention group post-test, 238.13 for the debate group pre-test, and 245.09 for the debate group post test.

Effect sizes for the ANOVA test was calculated because no other individual item or total score had statistically significant differences (Table 33). Eta-squared was calculated for how group membership impacts scores on the given item. The highest eta-squared was for argument construction, but still reflected a small effect size. And the small effect size was reported across each argument instrument item and for the total overall score. Additionally, Cohen's d was calculated for group comparisons on the post-test use of the argument education assessment instrument (Table 33). Only post-test was

used because hypothesis 2 was only concerned with comparison between groups rather than within a group. The effect sizes ranged from small to medium, with four comparisons indicating an effect size of zero. When interpreting the effect sizes it is important to refer back to the original means for each group because some of the comparisons reflect a decreased score rather than an increase in score.

Table 33

ANOVA group comparison effect sizes on the argumentation instrument

Groups compared by item	Eta-squared	P-value
Argument evaluation	0.028	0.14
Argument Utilization	0.032	0.11
Argument Bias	0.012	0.65
Argument Construction	0.033	0.04*
Argument Prioritization	0.027	0.15
Argument Total	0.010	0.72

* Significant between-subject ANOVA tests at the .05 level

Table 34
Group comparison effect sizes on the argumentation instrument post-test

Groups compared by item	Cohen's D	Magnitude
Argument evaluation		
Control vs Curricular 1	0.26	Small
Control vs Curricular 2	0.46	Medium
Control vs Debate	0.48	Medium
Curricular 1 vs Curricular 2	0.23	Small
Curricular 1 vs Debate	0.24	Small
Curricular 2 vs Debate	0.00	None
Argument utilization		
Control vs Curricular 1	0.47	Medium
Control vs Curricular 2	0.00	None
Control vs Debate	0.19	Small
Curricular 1 vs Curricular 2	0.38	Medium
Curricular 1 vs Debate	0.48	Medium
Curricular 2 vs Debate	0.21	Small
Argument bias		
Control vs Curricular 1	0.29	Small
Control vs Curricular 2	0.15	Small
Control vs Debate	0.16	Small
Curricular 1 vs Curricular 2	0.13	-
Curricular 1 vs Debate	0.13	-
Curricular 2 vs Debate	0.00	None
Argument construction		
Control vs Curricular 1	0.28	Small
Control vs Curricular 2	0.11	-
Control vs Debate	0.31	Small
Curricular 1 vs Curricular 2	0.40	Medium
Curricular 1 vs Debate	0.62	Medium
Curricular 2 vs Debate	0.20	Small
Argument prioritization		
Control vs Curricular 1	0.16	Small
Control vs Curricular 2	0.64	Medium
Control vs Debate	0.34	Small
Curricular 1 vs Curricular 2	0.46	Medium
Curricular 1 vs Debate	0.16	Small
Curricular 2 vs Debate	0.33	Small
Argument total		
Control vs Curricular 1	0.12	-
Control vs Curricular 2	0.00	None
Control vs Debate	0.24	Small
Curricular 1 vs Curricular 2	0.12	-
Curricular 1 vs Debate	0.35	Small
Curricular 2 vs Debate	0.22	Small

Study 4, spring 2017. In study 4, the changes adopted in study 3 continued – larger samples, an additional curricular intervention, and the use of pre-test. The scores for each group by item and overall total score are reported in Table 34. An ANOVA was conducted to compare the group scores (Control, Curricular 1, Curricular 2, and Debate) by item and overall total score. Levene's test was run to test for homogeneity of variances and found that the test was statistically significant for argument evaluation, argument utilization, and argument prioritization. Relying on Welch's F to compensate for the lack of homogeneity of variances, the adjusted F-ratio found statistically significant differences for argument utilization, *(F(7,52.899)=2.350, p.=.04,* η^2 *=.044).* Post hoc comparisons using the Tukey HSD test indicated the mean score for the control group pre-test (M=0.88, SD=0.52) was significantly different than the curricular 2 intervention post-test (M=1.11, SD=0.71). Also different were the control group post-test (M=0.71, SD=0.58) and curricular intervention 1 pre-test (M=0.78, SD=0.69) from the curricular 2 intervention post-test (M=1.11, SD=0.71). Unadjusted for homogenous variances, the ANOVA also found statistically significant differences for argument construction, *(F(7,401)=3.757, p.=.001,* η^2 *=.062),* and the overall argument instrument score, *(F(7,401)=3.088, p.=.004,* η^2 *=.051).* Tukey HSD post-hoc comparisons on argument construction found that the mean score for the control pre-test (M=1.48, SD=0.54) was different than both the curricular 1 post-test (M=1.12, SD=0.52) and the curricular 2 pre-test (M=1.49, SD=0.59). Furthermore the post-hoc comparison found the curricular 2 pre-test (M=1.49, SD=0.59) was statistically different from the curricular 1 post-test (M=1.12, SD=0.52).

A Kruskal-Wallis H test was run and showed that there were statistically significant difference in scores on the argument utilization item, argument prioritization item, argument construction item, and overall argument total score. For utilization, $x^2(7) = 16.220$, $p = .023$; with a mean rank of 217.10 for the control group pre-test, 187.94 for the control group post-test, 188.58 for the curricular 1 intervention group pre-test, 197.78 for the curricular 1 intervention group post-test, 223.82 for the curricular 2 intervention group pre-test, 250.21 for the curricular 2 intervention group post-test, 137.86 for the debate group pre-test, and 178.71 for the debate group post test. For prioritization, $x^2(7) = 15.317$, $p = .032$; with a mean rank of 215.30 for the control group pre-test, 211.30 for the control group post-test, 213.29 for the curricular 1 intervention group pre-test, 205.16 for the curricular 1 intervention group post-test, 200.61 for the curricular 2 intervention group pre-test, 193.19 for the curricular 2 intervention group post-test, 74.50 for the debate group pre-test, and 136.14 for the debate group post test. With argument construction, $x^2(7) = 29.565$, $p = .000$; with a mean rank of 239.25 for the control group pre-test, 201.89 for the control group post-test, 191.40 for the curricular 1 intervention group pre-test, 160.13 for the curricular 1 intervention group post-test, 241.62 for the curricular 2 intervention group pre-test, 204.86 for the curricular 2 intervention group post-test, 191.07 for the debate group pre-test, and 103.35 for the debate group post test. Finally, for the overall argument total score, , $x^2(7) = 18.873$, $p = .009$; with a mean rank of 235.05 for the control group pre-test, 207.31 for the control group post-test, 189.18 for the curricular 1 intervention group pre-test, 183.90 for the curricular 1 intervention group post-test, 214.59 for the curricular 2 intervention group pre-test, 217.81 for the curricular

Table 35
Scores on the argumentation instrument

Group	Evaluation	Utilization	Bias	Construction	Prioritization	Total
Control Pre (n=88) (M, SD)	1.22, 0.42	0.88, 0.52	1.26, 0.75	1.48, 0.54*	1.23, 0.92	6.06, 1.46
Control Post (n=82) (M, SD)	1.18, 0.43	0.71, 0.58	1.20, 0.76	1.30, 0.58	1.20, 0.91	5.59, 1.90
Curricular_1 Pre (n=77) (M, SD)	1.15, 0.43	0.73, 0.66*+	0.98, 0.83	1.27, 0.53	1.21, 0.92	5.34, 1.66
Curricular_1 Post (n=60) (M, SD)	1.13, 0.37	0.78, 0.69+	1.07, 0.83	1.12, 0.52*+	1.15, 0.90	5.24, 1.76
Curricular_2 Pre (n=49) (M, SD)	1.19, 0.49	0.92, 0.61	0.96, 0.80	1.49, 0.59*+	1.10, 0.95	5.66, 1.79
Curricular_2 Post (n=39) (M, SD)	1.09, 0.30	1.11, 0.71*+	1.14, 0.84	1.33, 0.52	1.06, 0.94	5.74, 1.74
Debate Pre (n=7) (M, SD)	1.00, 0.58	0.43, 0.53	1.43, 0.61	1.29, 0.57	0.00, 0.00	4.14, 1.35
Debate Post (n=7) (M, SD)	1.07, 0.61	0.71, 0.70	0.86, 0.69	0.86, 0.38	0.57, 0.79	4.07, 1.90
Total Pre (n=221) (M, SD)	1.18, 0.44	0.82, 0.60	1.10, 0.80	1.40, 0.56	1.16, 0.93	5.66, 1.65
Total Post (n=188) (M, SD)	1.14, 0.39	0.82, 0.66	1.13, 0.80	1.23, 0.55	1.13, 0.91	5.45, 1.84
Total (n=409) (M, SD)	1.16, 0.42	0.82, 0.63	1.11, 0.80	1.32, 0.56	1.15, 0.92	5.56, 1.74

* Significant between-subject ANOVA tests at the .05 level
+ Significant between-subject ANOVA tests at the .05 level

2 intervention group post-test, 105.14 for the debate group pre-test, and 116.43 for the debate group post test.

Effect sizes were calculated for the ANOVA test to provide practical significance (Table 35). The highest eta-squared was for argument construction, demonstrating that group membership could explain 6.2% of the variance in the argument construction score. Argument evaluation has the smallest eta-squared, with group membership accounting for only 1.1% of the variance in the item score fluctuation. Also, Cohen's d was calculated to better explain the practical difference between group scores on each argument item and the overall argument instrument score (Table 36). Again, only post-test scores were used for this effect size calculation. Cohen's d ranged from small to large, with all of argument evaluation showing a less than small magnitude of impact. The largest effect was between the curricular 2 intervention post-test and the debate post-test, with the curricular intervention performing over one standard deviation better than those students in the debate group. Again, this demonstrates the necessity for going back to the original means and standard deviations for interpreting the directionality of the magnitude.

Table 36
ANOVA group comparison effect sizes on the argumentation instrument

Groups compared by item	Eta-squared	P-value
Argument evaluation	0.011	0.695
Argument Utilization	0.044	0.036*
Argument Bias	0.024	0.194
Argument Construction	0.062	0.001*
Argument Prioritization	0.038	… **
Argument Total	0.051	0.004*

* Statistically significant at the .05 level
** Robust test of equality of means could not be performed because one group has 0 variance

Table 37
Group comparison effect sizes on the argumentation instrument post-test

Groups compared by item	Cohen's D	Magnitude
Argument evaluation		
Control vs Curricular 1	0.06	-
Control vs Curricular 2	0.11	-
Control vs Debate	0.12	-
Curricular 1 vs Curricular 2	0.12	-
Curricular 1 vs Debate	0.12	-
Curricular 2 vs Debate	0.04	-
Argument utilization		
Control vs Curricular 1	0.11	-
Control vs Curricular 2	0.62	Medium
Control vs Debate	0.00	None
Curricular 1 vs Curricular 2	0.47	Medium
Curricular 1 vs Debate	0.10	-
Curricular 2 vs Debate	0.57	Medium
Argument bias		
Control vs Curricular 1	0.16	Small
Control vs Curricular 2	0.07	Small
Control vs Debate	0.47	Medium
Curricular 1 vs Curricular 2	0.08	Small
Curricular 1 vs Debate	0.28	Small
Curricular 2 vs Debate	0.36	Small
Argument construction		
Control vs Curricular 1	0.33	Small
Control vs Curricular 2	0.05	-
Control vs Debate	0.90	Large
Curricular 1 vs Curricular 2	0.40	Medium
Curricular 1 vs Debate	0.57	Medium
Curricular 2 vs Debate	1.03	Large
Argument prioritization		
Control vs Curricular 1	0.06	-
Control vs Curricular 2	0.15	Small
Control vs Debate	0.74	Large
Curricular 1 vs Curricular 2	0.10	-
Curricular 1 vs Debate	0.72	Large
Curricular 2 vs Debate	0.59	Medium
Argument total		
Control vs Curricular 1	0.10	Small
Control vs Curricular 2	0.08	-
Control vs Debate	0.80	Large
Curricular 1 vs Curricular 2	0.29	Small
Curricular 1 vs Debate	0.64	Large
Curricular 2 vs Debate	0.92	Large

Hypothesis 3a and 3b: Comparing within group (varying levels of argument education curricular integration) pre/post scores on the argumentation education assessment instrument. Both versions of hypothesis 3 addressed the within group differences on argument item scores on the argumentation education assessment instrument. The within group differences between the pre-test and post-test scores were of particular interest. Only the fall 2016 and spring 2017 samples are analyzed here because they both administered the instrument as a pre-test and post-test. Further, only those participants who could be identified as having completed the pre-test and the post-test were considered for analysis. To answer these hypotheses, a paired samples t-test was run to test the within group differences on the five argument scale items.

Study 3, fall 2016. No participants in the control group could be identified as having completed both the pre- and post-test. As a result, only paired responses from the curricular 1 (n=66), curricular 2 (n=28), and debate group (n=11) were used. The scores for each group pre- and post-test are reported in Table 37 by item and total score. Paired samples t-tests were conducted to compare pre- and post- argument instrument item scores within each of the three groups. Only the argument construction item for the group curricular 1 yielded statistically significant results with the pre-test score (M=1.46, SD=0.50) higher than the post-test score (M=1.25, SD=-.54), $t(65)=2.96$, $p=.004$. Small sample sizes may have impacted the ability to identify statistically significant differences. Effect sizes were calculated in the form of a corrected Cohen's d for paired sample t-tests (Table 38) as they might help provide evidence to answer the hypotheses and provide researchers guidance moving forward. For example, none of the effect sizes had more than a small magnitude. The curricular 1 and debate group both had most of their small

effect sizes in a negative direction between pre-test and post-test while the curricular 2 group did show small practical increases on two of the argument instrument items between the pre- and post-test.

Table 38
Paired sample scores on the argumentation instrument fall 2016

Group	Evaluation	Utilization	Bias	Construction	Prioritization	Total
Curricular_1 Pre (n=66) (M, SD)	1.12, 0.41	0.85, 0.56	1.17, 0.76	1.46, 0.50*	1.44, 0.65	6.04, 1.71
Curricular_1 Post (n=66) (M, SD)	1.17, 0.45	0.70, 0.64	1.22, 0.73	1.25, 0.54*	1.47, 0.61	5.81, 1.60
Curricular_2 Pre (n=28) (M, SD)	1.09, 0.53	0.82, 0.58	1.29, 0.71	1.57, 0.45	1.38, 0.59	6.14, 1.76
Curricular_2 Post (n=28) (M, SD)	1.29, 0.62	1.00, 0.72	1.25, 0.70	1.48, 0.55	1.16, 0.64	6.18, 1.71
Debate Pre (n=11) (M, SD)	1.41, 0.49	1.18, 0.75	1.14, 0.78	1.64, 0.39	1.45, 0.52	6.82, 1.45
Debate Post (n=11) (M, SD)	1.32, 0.46	1.18, 0.75	1.05, 0.79	1.68, 0.40	1.45, 0.52	6.68, 1.03

* Significant between-subject ANOVA tests at the .05 level

Table 39
Paired sample t-test and effect size fall 2016

Group compared by item	Cohen's d	Magnitude
Curricular 1 (n=66)		
Argument evaluation	0.10	None
Argument utilization	0.22	Small
Argument bias	0.07	Small
Argument prioritization	0.04	Small
Argument construction*	0.34	Small
Curricular 2 (n=28)		
Argument evaluation	0.28	Small
Argument utilization	0.25	Small
Argument bias	0.07	None
Argument prioritization	0.27	Small
Argument construction	0.15	Small
Debate (n=11)		
Argument evaluation	0.17	Small
Argument utilization	0.00	None
Argument bias	0.14	None
Argument prioritization	0.00	None
Argument construction	0.07	None

* Statistically significant at the .05 level

Study 4, spring 2017. The scores for each paired sample group, by item and total score, are reported in Table 39. For this spring sample, the three groups with paired participants were curricular 1 (n=60), curricular 2 (n=34), and debate (n=4). Again, the control group was not able to produce any participant who completed the instrument as both a pre- and post-test. Paired sample t-tests were run on each of the five argument education assessment instrument items within each group, measuring the difference between the pre-test and post-test. Once more, the only statistically significant finding came from the Curricular 1 group on the argument construction item with the pre-test (M=1.33, SD=0.38) scoring higher than the post-test (M=1.17, SD=0.42), $t(59)=2.21$, $p=.03$.

Again, the effect sizes for each of the item scores were run to be able to examine the practical significance even in the case of no statistical significance. The effect size and whether or not the paired-sample t-test was statistically significant are reported in Table 39. All paired sample t-test results are reported, not just the ones found statistically significant because sample size concerns may impact the reported p-values. Recall that to properly interpret the effect size, it is important to refer back to the original group means in Table 39 to understand the directionality of the effect. For example, the curricular 1 intervention group actually decreases scores between the pre- and post-test in three of the five items while the debate extra-curricular group increases in three of the five items between the pre- and post-test. Generally the effect sizes are fairly small across the items with the exception of the debate group where three of the five items have a large effect size.

Table 40
Paired sample scores on the argumentation instrument spring 2017

Group	Evaluation	Utilization	Bias	Construction	Prioritization	Total
Curricular_1 Pre (n=60) (M, SD)	1.16, 0.30	0.75, 0.54	0.98, 0.62	1.33, 0.38*	1.20, 0.68	5.42, 1.48
Curricular_1 Post (n=60) (M, SD)	1.14, 0.20	0.87, 0.49	1.10, 0.63	1.17, 0.42*	1.12, 0.66	5.39, 1.29
Curricular_2 Pre (n=34) (M, SD)	1.19, 0.43	0.82, 0.82	1.08, 0.67	1.38, 0.41	1.09, 0.57	5.57, 1.45
Curricular_2 Post (n=34) (M, SD)	1.16, 0.27	1.09, 0.96	1.19, 0.62	1.32, 0.39	1.01, 0.78	5.65, 1.22
Debate Pre (n=4) (M, SD)	0.88, 0.25	0.25, 0.50	1.13, 0.75	1.63, 0.48	0.50, 0.58	4.38, 2.17
Debate Post (n=4) (M, SD)	1.50, 0.41	0.75, 0.29	1.25, 0.65	1.25, 0.29	0.50, 0.71	5.25, 0.87

* Significant between-subject ANOVA tests at the .05 level

Table 41
Paired sample t-test and effect size spring 2017

Group compared by item	Cohen's d	Magnitude
Curricular 1 (n=60)		
Argument evaluation	0.04	None
Argument utilization	0.19	Small
Argument bias	0.17	Small
Argument prioritization	0.15	Small
Argument construction	0.28	Small
Curricular 2 (n=34)		
Argument evaluation	0.07	None
Argument utilization	0.23	Small
Argument bias	0.11	None
Argument prioritization	0.08	None
Argument construction	0.10	None
Debate (n=4)		
Argument evaluation	0.83	Large
Argument utilization	0.63	Large
Argument bias	0.11	None
Argument prioritization	0.00	None
Argument construction	0.57	Large

CHAPTER 5

Discussion

It is increasingly apparent and understood by researchers and the public alike that an important skill of leaders is the ability to formulate and evaluate arguments. Arguments support or refute decisions that affect all of society, and better leaders must be trained in argument education. Within the United States, colleges and universities can perform a critical role in leadership development. Enrollment and graduation statistics show that a significant portion of Americans attend and eventually graduate from institutions of higher learning. But what are the pedagogical practices that can help students develop into being transformational leaders? And how are we supposed to know when these practices are effective? The research conducted here was concerned with argument education as one way that postsecondary education could demonstrate the value of a college degree. However, the ability to define, measure, and demonstrate what constitute argument education was missing in current approaches to argumentation across the curriculum. Specifically, the studies completed here move forward a research agenda toward developing and validating an argument education assessment instrument that could be utilized for learning and assessing argument education across the higher education curriculum.

The results for the three studies did not support Hypothesis 1 that the instrument would yield a g-coefficient at or above 0.70. In fact, the g-coefficient steadily decreased over the three semesters, from 0.43 in spring 2016 to 0.23 in spring 2017. Hypothesis 1 emphasized the g-coefficient for inter-rater reliability because of its ability to partition out error rather than treating it as de-differentiated. Looking at the actual sources of

variance for the person/rater*item design highlights that the variance is largely due to the item and item interactions, not the raters. The high item variance was true across each of the three studies. In particular, the person by item variance accounted for the highest percentage of variance. In study 2, person by item variance was 74%. In study 3 and study 4, it was 72% and 85% respectively. With the low g-coefficient and simultaneous low rater variance, it was important to look at the other measures of inter-rater reliability for additional context.

The consensus and consistency estimates of inter-rater reliability suggest that the argument education assessment instrument may still be considered reliable. The different coefficients for consistency and consensus all approach or far exceed the 0.70 benchmark. Argument construction in study 2 and study 3 is the exception, ranging from 0.47 to 0.72. Cohen's Kappa, the often-cited consensus estimate of inter-rater reliability, was well above 0.70 for three of the five items in the most recent study. But even the different iterations of Kappa ranged from right around 0.70 to as high as .080 and 0.95 across all five items in the final study.

The Cronbach's alpha results confirm that it is the items and not the raters that are not consistent. Cronbach's alpha reports how well the items group together internally within an instrument or scale. For each of the three studies, Cronbach's alpha was well below the suggested level of 0.70. The alpha ranged from .425 in the first study in spring 2016 to 0.272 in the third study in spring 2017. Furthermore, the inter-item correlations between the five items on the instrument were also low across all three studies.

Hypothesis 1 was not supported, but the results suggest that the evidence can still confirm the reliability of the argument education assessment instrument. For reliability in

this performance-based testing situation, the agreement between raters was of paramount interest. Of course, reliability is not an all or nothing thing. One cannot definitively say that a given set of evidence concludes an instrument is reliable. But given that the results suggested the raters did consistently agree on the observed behaviors across items and all three studies, the instrument can be treated as reliable for these learning and testing environments. And the results suggest that further research is needed into the reliability of this instrument.

The results comparing scores on the argument education assessment instrument between groups only partially supported Hypothesis 2a and 2b. For 2a, it was hypothesized that students with exposure to curricular argument education interventions would report higher scores on the instrument than students from a control group. Across all three studies, the results did not completely support this. First, only a few of the items on some of the groups showed statistically significant differences across the three studies. And in some instances, those were differences were in the opposite direction of what would have been expected. While not statistically significant, study 3 did report some small to medium magnitude effect size differences between the control and two curricular groups on some of the argument items. These results suggest that the instrument may be able to detect small differences between the argument education curricular intervention and control groups, but the evidence is not overwhelming.

The data partially supported Hypothesis 2b, that members of the debate extra-curricular group would report higher scores on the assessment instrument than either the control or curriculum groups. Again, the ANOVA from study 2 did show statistically significant differences between the debate group and both control and curricular groups

on the argument evaluation item. But study 3 and study 4 showed no statistically significant differences. Given the small sample sizes for the debate group, it may been difficult to detect statistically significant differences. For study 2, the debate group did have moderate to large effect sizes in their respective higher differences on the argument education items. But this study only had six participants in the debate group, while there were over forty in each of the other groups. Study 3 found the debate group higher on some items, like evaluation, than control and curricular 1, but about the same on items like construction. Effect sizes here were small to moderate when the debate group reported higher scores. The results do not call for a complete rejection of Hypothesis 2b, but suggest that additional studies are needed to confirm the identified group differences here.

A couple of reasons exist that might help explain the lack of support for hypothesis 2a and 2b. First, the control group may not have been as much of a control group as originally thought. While the control group is selected from an entry-level general education course, the course curriculum does include some instructional elements that address argument education. Second, implementation fidelity is another concern for both the control and curriculum groups. For the control groups, faculty may do very little with the embedded argument education elements or a given faculty member may center their course around the argument elements. On the curriculum side, there were no checks to ensure that the course content and instruct aligned closely with the argumentation construct categories. Some of the activities, like classroom debates, include elements like asking a student to construct an argument. But the faculty member may not devote classroom time to teaching the students how to actually construct a sound argument.

These issues combined may help explain why the curricular and control groups are sometimes closer together in their scores on the instrument. Second, the unequal sample sizes may impact the ability to detect statistically significant differences between groups. In study 2, the control and curricular groups had roughly equal sizes, but study 3 and study 4 found groups ranging from low 30s to right around 90. Furthermore, the debate group sample sizes were extremely small in comparison, ranging from 6 to 19 at its highest. As a result of this thinking, the practical significance was also used to address the hypothesis. And in identifying some practical effect sizes, it is believed that the argument education instrument is able to identify some meaningful differences between groups even if the hypotheses are not totally supported by the results.

Hypothesis 3a and 3b regarding within group pre- and post-test increases on the argument education instrument were both partially supported by the results. Hypothesis 3a addressed the within group scores for participants in the curricular intervention groups. Both study 3 and study 4 found practically significant increases within curricular 1 and curricular 2 groups on the argument evaluation, utilization, and bias items. The practical significance for these within group increases was of a small magnitude. Interestingly, curricular 1 and 2 were not consistent where they increased. For example, curricular intervention 1 saw an increase on bias, but a decrease on utilization. While curricular group 2 saw the inverse on the directionality of their utilization and bias item scores. Argument construction saw a statistically significant decrease, possibly reflecting an issue with the item as is discussed among the limitations of the study.

The results only partially supported hypothesis 3b concerning the within group pre/post increase in argument item scores. Study 3 found no statistically significant

evidence to support this hypothesis. The debate group here did not identify any statistically or practically significant increases in their pre/post argument education scores. However, like Hypothesis 3a, study 4 did provide evidence to at least partially support Hypothesis 3b. There was practically significant evidence suggesting that on argument evaluation, utilization, and bias that debate students did score higher after being exposed to the extra-curricular debate intervention. But the difference here was that the debate group saw increases that had effect sizes of a large magnitude. Given the small sample size for study 4 (n=4) and the increases were not across the board, it would be difficult to conclude these studies provide overwhelming support for Hypothesis 3b. But at the very least, the argument education assessment instrument was able to identify some within group increases on the instrument after exposure to some form of curricular intervention.

Limitations

As with any research study, the research conducted here also contained limitations. Some of the limitations have already been mentioned, for example concerns about sample size and implantation fidelity. Another limitation for the studies was the reliance on samples of convenience. Students were not randomly assigned into groups. Instead each participant was recruited or participated because the class was offered at the researcher's home institution or was an argumentation or debate colleague.

Generalizability presents another limitation to this research. While some of the participants are from different institutions and represent different backgrounds, most of the participants come from one academic institution. And the student body of that institution is not representative of the large student body in U.S. postsecondary education.

Instrument administration also presented a limitation. The same argument education items were given at the beginning and end of the semester. Seeing the same item again may have impacted the participants' ability to respond, positively or negatively. Motivation to complete the survey was also an administration limitation. Students at the end of the semester may have experienced fatigue, lack of motivation, or just not care anymore about completing a research survey beyond the scope of the course.

The raters might have been another limit to this study. Only two individuals rated items throughout the duration of the research. Additionally, it was the same two raters the entire time. And one of the raters was the author of this study. Perhaps increasing the number of raters or diversifying the raters could have positively impacted the inter-rater reliability coefficients.

The assessment instrument prompts themselves could have also served as a limitation to the study. While the data from the subject matter experts suggested that the prompts were appropriate and aligned with measuring argument education, perhaps other concerns existed. The subject matter experts did not evaluate the prompts for research participant accessibility. Perhaps one or more of the prompts were not accessible to the research sample or could have initiated a triggering effect. The argument construction prompt, for example, asked respondents to make an argument for and against the death penalty. Maybe the content areas of this or other prompts were too close to a student's experience or could not solicit an adequate response to rate because of lack of content familiarity.

A final major limitation was the lack of control over curricular content. And this was true across all of the groups. Lack of control over curricular content made it hard to

ensure the intended consistency and fidelity to the differentiated group. For example, the control should have had minimal argument education. Or when consulting with faculty to implement argument education in their class, the inability to ensure it is implement as designed.

Implications

The research conducted here has several implications for those interested in studying argument education and assessment. First, the lack of consistency or reliability among argumentation items suggests revisiting the defined construction of argumentation for education and assessment purposes. While the subject matter experts seemed to agree with how the construct was defined, operationalized, and assessed – the data did not seem to align with this understanding. One possible exception is argument construction, with both the experts and research data pointing to its problematic inclusion. Here, the subject matter experts disagreed some about its role in defining argument as a construct. Moreover, one expert commented that the prompt for construction did need more detail provided. The inter-rater reliability data also revealed argument construction as the item producing the least consistent scores.

The high item by person interaction from the g-study and low Cronbach's alpha provided evidence that the items may not fit within one unified construct. They could be two or even up to five different constructs at play within the field of argumentation. Further empirical research and follow-up with subject matter experts is needed to test argumentation as a unified construct, a series of skills, or something else. Second, and possibly related to the first, argumentation as a construct needs further distinguishing from critical thinking. More research is needed at a theoretical and empirical level to

understand how critical thinking and argumentation relate to one another. For example, one could administer both a critical thinking instrument alongside the argument education instrument to see how the scores correlate with one another. Perhaps sending the same (or revised) survey to critical thinking subject matter experts would be another way to tease out the differences and similarities between argumentation and critical thinking.

Third, the actual constructed responses themselves could be analyzed for additional information about argumentation. This would be a labor-intensive process as there were nearly 1,000 participants and five items per person. But in rating the responses, some themes emerged anecdotally. For example, in looking at argumentation bias responses, the notion of bias as present when only one side of an issue is presented came up repeatedly. This is counter to how bias is defined and operationalized for the study, but may be worth research more and even integrating more explicitly into argument instruction. Another example came up in argument evaluation. A number of respondents were evaluating the better argument based on which response used more manipulative rhetoric (hiding the brutality of testing on animals) and not on whether or not the argument explicitly contained the reasoning linking the argument together. An implication of this is that we may change how we understand or at least teach argumentation.

Fourth, and perhaps most importantly, more research is needed on assessing argument education. The studies conducted here did provide evidence to suggest the beginning validation of the proposed argument education assessment instrument. But this validation process is by no means over. And other interested scholars and practitioners of argument should be interested in more direct measures of learning, teaching, and

practicing argument. For example, additional items could be piloted to test the item interaction. The original instrument had two scenario prompts drafted for each item. They could both be piloted simultaneously to see if the item interaction is unique to the item, prompt, or the proposed construct. Furthermore, more diverse and larger samples are needed to improve the generalizability and power of the studies. The debate sample contains the potential to have the largest differences, but has been impossible to detect statistically significant differences with 6-17 participants in a given sample.

Fifth, perhaps the results of the inter-rater reliability analyses have implications for how generalizability is theorized and practiced. The g-coefficients here were low, but not necessarily a reason to evaluate the instrument as not reliable. The g- and d-studies functioned more as a useful diagnostic tool to help provide more evidence-based context for interpreting the reliability of the instrument's scores. When taken as one piece of the puzzle along side other estimates of inter-rater reliability, g-theory can provide a better picture of how different elements or faces of a researcher's design are impacting the scores.

Sixth, and finally, the study highlights some of the potential for including argument education into leadership development. For example, the argument skills to utilize argument in a given situation or prioritize argument for an audience are essential skills for transformational leadership (Buller, 2014). The ability to read an organizational environment and then build a case for change for that situation is a hallmark of organizational leadership. Other skills like argument construction and evaluation are essential for leaders to be able to participate in meaningful decision-making and deliberations that effects the long-term strategic planning of an organization (Eckel &

Kezar, 2003). Here, leaders are able to utilize argument evaluation and construction skills to gauge a situation and then develop persuasive cases for impacting the underlying structures or values of a specific organization.

Conclusion

The current research study contributed to the literature on leadership and argument education within higher education. Specifically, the study here suggested that argument education assessment is possible but additional is needed in how we define argumentation such that it may be learned, practiced, and assessed. This is especially important as American post-secondary educational institutions are facing mounting pressures to justify the value of (and significant investment in) a college education. This study makes the case that argument education is key to developing transformational leaders. Preparing generations of leaders can be one way to demonstrate the worth of an investment in higher education, whether from a family, institution, state, or public's perspective. However what was missing in the current research on postsecondary argument education were ways to more directly measure if argumentation skills were being learned, developed, and/or practiced by students. Building on existing approaches to argument education, the present study developed argumentation as a construct for the purposes of teaching and measuring argument education. In particular, the research here tested an argumentation education assessment instrument over the course of three academic semesters and multiple semesters that could be used to assess argument education across higher education.

APPENDIX

Appendix A
Argumentation Education Assessment Instrument Rubric

	Unsatisfactory		Fair		Good
Identify biased argument (Bias is recognized as preconceived opinions without supporting reasons for one's position)	Does not identify any bias in claims.		Identifies the presence of bias in an attempt to induce change but does not explain the appropriate source of or motivation for bias.		Identifies the presence of bias and explains the appropriate source of or motivation for bias given the specific attempt to create change.
	Unsatisfactory		Fair		Good
Prioritize information based on situation	Does not attempt to distinguish between relevant and irrelevant information for the situation.		Distinguishes between relevant and irrelevant information but does not prioritize appropriately for the situation.		Distinguishes between relevant and irrelevant information and prioritizes appropriately for the situtaion.
	Unsatisfactory		Fair		Good

Logical argument construction	Argument contains one element (claim, warrant, data) of a logical argument but does not present a complete argument.		Argument contains two elements (claim, warrant, data) of a logical argument but does not present a complete argument.		Constructs a complete logical argument, including claim, warrant, and data.
	Unsatisfactory		Fair		Good
Argument evaluation	Does not evaluate any arguments present.		Attempts to evaluate arguments but does not identify the relevant & sufficient reasoning based on the perspective established in the argument.		Evaluates argument by identifying the relevant & sufficient reasoning based on the perspective established in the argument.
	Unsatisfactory		Fair		Good
Argument utilization in a situation	Does not fit arguments for the situation.		Attempts to fit arguments for the given situation but misjudges the situation, purpose, or audience.		Fits the arguments for the situation, purpose, and audience.
	Unsatisfactory		Fair		Good

Identify affective argument (Affect is understood as emotions, feelings, attitudes, values, or other relational dimensions of argument)	Does not identify any affective dimension of argument at play; whether audience, advocate, or situation.		Identifies an affective dimension of at least one perspective but the affective argument identified is not relevant for the given situation.		Identifies an affective dimension of at least one perspective and the affective argument identified is relevant for the given situation.

Appendix B
Argumentation Education Assessment Instrument Prompts

Identify biased argument
The National Collegiate Athletic Association (NCAA) is considering multiple proposals to begin paying student-athletes because currently student-athletes receive no compensation for their athletic play. The NCAA is a non-profit organization that governs all athletes and athletic programs that compete in intercollegiate athletics. For example, they make sure student-athletes are academically eligible and that athletic programs are following NCAA guidelines for competition. In addition to governing student-athletes, the NCAA also receives billions of dollars annually from corporate sponsorship for their major sporting events like the football playoffs and March basketball tournaments.
The NCAA has recently issued a press release arguing against the proposals to begin paying student-athletes. In the press release, they present several reasons why student-athletes should not be paid. First, the NCAA suggested that paying student-athletes would transform them from students first and athletes second to athletes first and students second. Second, paying student-athletes would only benefit the largest schools and most popular sports. Finally, paying student athletes would increase the competition and reward for athletic participation. This would create an environment that would encourage more cheating, use of performance enhancing drugs, and other scandals the NCAA hopes to avoid. Is there any bias present in the NCAA press release against paying student-athletes? Explain why or why not.

Prioritize information based on situation
A local district school board is considering a petition by some families within the school district to make the school start times later. The school board members are charged with ensuring a quality education for all students and keeping down the costs of education for families in the school district. The families have asked you to help them construct their case for why school start times should be later for elementary, middle, and high school students. How would you construct the case to be presented in front of the school board for later school start times?

Argument construction
Construct a position for and a position against a Federal law banning the use of the death penalty in the United States.

Argument evaluation
Identify the better argument, below, for the proposition "Animal testing is justified." Please explain why your selection is the better argument.
1. Animal testing is justified because it saves human lives to test products and treatments on animals first before using the products and treatments on humans.
2. Animal testing is justified because it saves human lives due to the countless medical breakthroughs that happen every year.

Argumentation – argument utilization in a situation

You are the hiring manager on a job search for a new administrative assistant in your company. Select one of the two final candidates for this position and justify why you decided to hire them.

The mission of your company is to provide excellent customer service in connecting local businesses with temporary contract workers. Your company has been operating successfully within the area for over fifty years. Just this year, the company is celebrating its 2nd straight award for customer service. The administrative assistant position will provide administrative support for the office manager. The administrative assistant will support the office manager by doing handling all office communication (telephone, front office, company email) conducting background checks on potential contract workers, and advertising positions to local businesses. This position is new because of the company's success and growth within the local business area. In fact, the profits grew 25% over the last year.

Candidate A: 10 years of customer service experience working for a bank. During the interview, candidate A demonstrated they had done prior research on your company, including mentioning the awards for customer service. Additionally, candidate A worked for two years in college as a contract worker. Finally, candidate A has three years of experience writing stories for the local newspaper.

Candidate B: 10 years of administrative experience working for a bank. During the interview, candidate B demonstrated they had outstanding interpersonal skills. Furthermore, candidate B has worked as a marketing intern in college and included some advertising examples with their resume. Finally, candidate B has three years of experience supervising other employees.

Identify affective argument

A local professional sports stadium has decided to eliminate all of the unhealthy foods and drinks from its stadium restaurants, catering, concession stands, and neighborhood eating establishments near the stadium. They decided to remove foods that do not meet government standards because the average citizen was gaining too much weight. As a city and state funded stadium, they believed that eliminating these foods was part of their mission to provide a healthy and safe entertainment environment. The local merchants who provided the now banned food and drinks have a meeting with the stadium owners to try and reverse their decision. Pretend you are one of the local merchants, how do you construct your case so that the stadium will again sell your food and drinks?

Appendix C
Argument as Critical Thinking Pilot Validity Survey Sp16

The purpose of this research is to investigate what impact, if any, argument education has on a student's argumentation skills. The larger context for this project is my interest in generating empirical research on the educational impact of debate pedagogy and argument education. I was not satisfied with the current assessment instruments on argumentation or critical thinking; they were absent, not applicable, or resource intensive. This survey is intended to help evaluate and develop a valid instrument that can be used for assessing argument as critical thinking.

David Zarefsky's definition of argumentation in the 2001 Encyclopedia of Rhetoric is the definition used in this study to understand and operationalize argument as critical thinking. His definition states, "Argumentation is the study of reason-giving used by people to justify their beliefs and values and to influence the thought and action of others. Its central concern is with the rationality or reasonableness of claims put forward in discourse. This, in turn, depends on whether the claims are warranted, or grounded in evidence and inference that are themselves acceptable and hence constitute good reasons for the claim." Do you agree this is an acceptable foundational definition of argumentation?

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

Is there anything you would include or exclude in this definition that is not already stated?

For this study, I needed to operationalize argumentation into observable behaviors so that one could evaluate whether or not a student has learned argumentation skills. I have identified five performance criteria as essential to foundational argumentation skills. These criteria are a synthesis of more specific behaviors from the 1990 Jones et al "National Assessment of College Student learning: Identifying College Graduates' Essential Skills in Writing, Speech and Listening, and Critical Thinking." The Jones et al study triangulates agreement on specific skills between faculty, employers, and policymakers. I went through and identified the critical thinking behaviors that actually seemed like essential skills in argumentation. For each performance criteria (and subsequent behaviors), please rate whether you agree or disagree that it is a foundational argumentation skill.

Identify biased argument (recognize use of misleading language, recognize use of slanted definitions/comparisons, determine if an argument rests on false, biased, or doubtful assumptions) is a foundational argumentation skill.

- ❍ Strongly agree (1)
- ❍ Agree (2)

- Neither agree nor disagree (3)
- Disagree (4)
- Strongly disagree (5)

Prioritize information based on situation (detect introduction of irrelevant information into an argument, recognize relationship between communication purpose and ideas that must be resolved to achieve this purpose, identify background information provided to explain reasons which support a conclusion, assess the importance of an argument and determine if it merits attention, judge what background information would be useful to have when attempting to develop a persuasive argument in support of one's opinion) is a foundational argumentation skill.

- Strongly agree (1)
- Agree (2)
- Neither agree nor disagree (3)
- Disagree (4)
- Strongly disagree (5)

Logical argument construction (identify the unstated assumptions of an argument, determine if one has sufficient evidence to form a conclusion, present an argument succinctly in such a way as to convey the crucial point of an issue, cite relevant evidence and experiences to support their position, seek various independent sources of evidence, rather than a single source of evidence, to provide support for a conclusion) is a foundational argumentation skill.

- Strongly agree (1)
- Agree (2)
- Neither agree nor disagree (3)
- Disagree (4)
- Strongly disagree (5)

Argument evaluation (evaluate an argument in terms of its reasonability and practicality; evaluate the credibility, accuracy, and reliability of sources of information; assess statistical information used as evidence to support an argument; assess how well an argument anticipates possible objections, offers, when appropriate, alternative positions; determine and evaluate the strength of an analogy used to warrant a claim or conclusion; determine if conclusions based on empirical observations were derived from a sufficiently large and representative sample) is a foundational argumentation skill.

- Strongly agree (1)
- Agree (2)
- Neither agree nor disagree (3)
- Disagree (4)
- Strongly disagree (5)

Utilize argument in a situation (present supporting reasons and evidence for their conclusion(s) which address the concerns of the audience, develop and use criteria for making judgments that are reliable, intellectually strong and relevant to the situation at hand) is a foundational argumentation skill.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

Would you want any of the five identified performance criteria removed from a foundational understanding of essential argument skills? Please check all that apply.

- ❑ Identify biased argument (1)
- ❑ Prioritize information based on situation (2)
- ❑ Logical argument construction (3)
- ❑ Argument evaluation (4)
- ❑ Argument utilization in a situation (5)
- ❑ None, I would remove none of the five behaviors. (6)

Are there other performance criteria you would want added as essential to argumentation skills?

Below are sections of a rubric created based on the performance criteria considered essential for foundational argumentation skills. This rubric would be used to rate student generated responses assessing whether or not they have demonstrated argument as critical thinking. Please rate whether or not the levels for the performance criteria are clear and reflective of the solicited performance.

	Unsatisfactory	Fair	Good
Identify biased argument (Bias is recognized as preconceived opinions without supporting reasons for one's position)	Does not identify any bias in claims.	Identifies the presence of bias but does not explain the appropriate source of or motivation for bias.	Identifies the presence of bias and explains the appropriate source of or motivation for bias.
	Unsatisfactory	Fair	Good
Prioritize information based on situation	Does not attempt to distinguish between relevant and irrelevant information for the situation.	Distinguishes between relevant and irrelevant information but does not prioritize appropriately for the situation.	Distinguishes between relevant and irrelevant information and prioritizes appropriately for the situtaion.
	Unsatisfactory	Fair	Good
Logical argument construction	Argument contains one element (claim, warrant, data) of a logical argument but does not present a complete argument.	Argument contains two elements (claim, warrant, data) of a logical argument but does not present a complete argument.	Constructs a complete logical argument, including claim, warrant, and data.
	Unsatisfactory	Fair	Good
Argument evaluation	Does not evaluate any arguments present.	Attempts to evaluate arguments but does not identify critical missing components or judge argument using appropriate criteria.	Evaluates argument by identifying critical missing components and judging argument using appropriate criteria.
	Unsatisfactory	Fair	Good
Argument utilization in a situation	Does not fit arguments for the situation.	Attempts to fit arguments for the given situation but misjudges the situation, purpose, or audience.	Fits the arguments for the situation, purpose, and audience.

The identify biased argument performance criteria rubric is clear and reflective of the solicited performance.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

The prioritize information based on situation performance criteria rubric is clear and reflective of the solicited performance.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

The logical argument construction performance criteria rubric is clear and reflective of the solicited performance.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

The argument evaluation performance criteria rubric is clear and reflective of the solicited performance.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

The argument utilization in a situation performance criteria rubric is clear and reflective of the solicited performance.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

Do you have any other feedback on the argument as critical thinking rubric?

To assess argument as critical thinking, I generated scenarios to solicit written responses from students where a rater could identify, through the use of the rubric, the behaviors associated with the argumentation performance criteria. Please review the scenarios and

indicate whether you agree or not that the scenario aligns with the given performance criteria.

Scenario 1: Identify the better argument, below, for the proposition "Animal testing is justified." Please explain why your selection is the better argument. 1. Animal testing is justified because it saves human lives to test products and treatments on animals first before using the products and treatments on humans. 2. Animal testing is justified because it saves human lives due to the countless medical breakthroughs that happen every year. Scenario 1 aligns with argument evaluation.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

Scenario 2: A local district school board is considering a petition by some families within the school district to make the school start times later. The school board members are charged with ensuring a quality education for all students and keeping down the costs of education for families in the school district. The families have asked you to help them construct their case for why school start times should be later for elementary, middle, and high school students. How would you construct the case to be presented in front of the school board for later school start times? Scenario 2 aligns with argument utilization.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

Scenario 3: The National Collegiate Athletic Association (NCAA) is considering multiple proposals to begin paying student-athletes because currently student-athletes receive no compensation for their athletic play. The NCAA is a non-profit organization that governs all athletes and athletic programs that compete in intercollegiate athletics. For example, they make sure student-athletes are academically eligible and that athletic programs are following NCAA guidelines for competition. In addition to governing student-athletes, the NCAA also receives billions of dollars annually from corporate sponsorship for their major sporting events like the football playoffs and March basketball tournaments. The NCAA has recently issued a press release arguing against the proposals to begin paying student-athletes. In the press release, they present several reasons why student-athletes should not be paid. First, the NCAA suggested that paying student-athletes would transform them from students first and athletes second to athletes first and students second. Second, paying student-athletes would only benefit the largest schools and most popular sports. Finally, paying student athletes would increase the competition and reward for athletic participation. This would create an environment that would encourage more cheating, use of performance enhancing drugs, and other scandals the NCAA hopes to avoid. Is there any bias present

in the NCAA press release against paying student-athletes? Explain why or why not. Scenario 3 aligns with identify biased argument.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

Scenario 4: Construct a position for and a position against a Federal law banning the use of the death penalty in the United States. Scenario 4 aligns with argument construction.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

Scenario 5: You are the hiring manager on a job search for a new administrative assistant in your company. Select one of the two final candidates for this position and justify why you decided to hire them. The mission of your company is to provide excellent customer service in connecting local businesses with temporary contract workers. Your company has been operating successfully within the area for over fifty years. Just this year, the company is celebrating its 2nd straight award for customer service. The administrative assistant position will provide administrative support for the office manager. The administrative assistant will support the office manager by handling all office communication (telephone, front office, company email), conducting background checks on potential contract workers, and advertising positions to local businesses. This position is new because of the company's success and growth within the local business area. In fact, the profits grew 25% over the last year. Candidate A: 10 years of customer service experience working for a bank. During the interview, candidate A demonstrated they had done prior research on your company, including mentioning the awards for customer service. Additionally, candidate A worked for two years in college as a contract worker. Finally, candidate A has three years of experience writing stories for the local newspaper. Candidate B: 10 years of administrative experience working for a bank. During the interview, candidate B demonstrated they had outstanding interpersonal skills. Furthermore, candidate B has worked as a marketing intern in college and included some advertising examples with their resume. Finally, candidate B has three years of experience supervising other employees. Scenario 5 aligns with argument utilization in a situation.

- ❍ Strongly agree (1)
- ❍ Agree (2)
- ❍ Neither agree nor disagree (3)
- ❍ Disagree (4)
- ❍ Strongly disagree (5)

Do you have any other feedback on the scenario prompts?

Thank you so much for taking the time to review and provide feedback on the first draft of this instrument to assess argument as critical thinking. If you have any additional feedback, please feel free to provide in the text box below or contact Paul Mabrey directly. Thank you!!!

References

Alkharusi, H. A. (2012). Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education*, *2*(1), 184-196.

Andrews, R. (2009a). *Argumentation in Higher Education: Improving practice through theory and research*. Routledge.

Andrews, R. (2009b). A case study of argumentation at undergraduate level in History. *Argumentation*, *23*(4), 547-558.

Andrews, R. (1995). *Teaching and learning argument*. Continuum Intl Pub Group.

Antaonakis, J. (2012). Tranformational and Charismatic Leadership. In D.V. Day & J. Antonakis (Eds.), *The nature of leadership* (256-288). Los Angeles, CA: Sage.

Argument Centered Education (no date). Argument and the Common Core. Retrieved from http://argumentcenterededucation.com/argument/argument-and-the-common-core/.

Bathgate, M., Crowell, A., Schunn, C., Cannady, M., & Dorph, R. (2015). The learning benefits of being willing and able to engage in scientific argumentation. *International Journal of Science Education*, *37*(10), 1590-1612.

Berkowitz, S. J. (2006). Developing critical thinking through forensics and communication education: Assessing the impact through meta-analysis. *Classroom communication and instructional processes: Advances through meta-analysis*, 43-59.

Berrett, D. (2013, September 18). Employers and Public Favor Graduates Who Can

Communicate, Survey Finds. *The Chronicle of Higher Education*. Retrieved from http://www.chronicle.com/article/EmployersPublic-Favor/141679/.

Buller, J. L. (2014). *Change leadership in higher education: A practical guide to academic transformation*. San Francisco, CA: John Wiley & Sons.

Camp, J. M., & Schnader, A. L. (2010). Using debate to enhance critical thinking in the accounting classroom: The Sarbanes-Oxley Act and US tax policy. *Issues in accounting education*, *25*(4), 655-675.

Christie, L. (2014, July 2). Employers value skills over college degrees, workers say. *CNN Money*. Retrieved from http://money.cnn.com/2014/07/02/pf/worker-skills/.

Cook, N. (2015, June 18). When It Comes to Getting a Job, Americans Believe Skills Trump College. *The Atlantic*. Retrieved from http://www.theatlantic.com/business/archive/2015/06/millennials-skills-college-career-success/395996/.

Darby, M. (2006). Debate: a teaching-learning strategy for developing competence in communication and critical thinking. *Journal of Dental Hygiene: JDH/American Dental Hygienists' Association*, *81*(4), 78-78.

Davidson, K. (2016, August 30). Employers Find ‘Soft Skills’ Like Critical Thinking in Short Supply. *The Wall Street Journal*. Retrieved from http://www.wsj.com/articles/employers-find-soft-skills-like-critical-thinking-in-short-supply-1472549400.

Deane, P., & Song, Y. (2015). The Key Practice, Discuss and Debate Ideas: Conceptual Framework, Literature Review, and Provisional Learning Progressions for Argumentation. *ETS Research Report Series*, *2015*(2), 1-21.

Deards, P. (2014, August 12). Making the case for teaching students to debate. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2014/08/12/01deards.h34.html.

Eckel, P. D., & Kezar, A. J. (2003). *Taking the reins: Institutional transformation in higher education*. Westport, CT: Praeger Publishers.

Ennis, R.H. and Weir, E. (1985). The Ennis-Weir critical thinking essay test. Pacific Grove, CA: Midwest Publications.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Los Angeles, CA: Sage Publishing.

Gallo, C. (2013, December 27). One Skill That Will Boost Your Value by Fifty Percent in 2014. *Forbes*. Retrieved from http://www.forbes.com/sites/carminegallo/2013/12/27/one-skill-that-will-boost-your-value-by-fifty-percent-in-2014/#4453e9831eab.

Gregory, M., & Holloway, M. (2005). The debate as a pedagogic tool in social policy for social work students. Social Work Education, 24(6), 617-637.

Hamilton, R. (2010, September 28). Data-Driven Accountability Emphasized in Higher Ed. *The Texas Tribune*. Retrieved from https://www.texastribune.org/2010/09/28/data-driven-accountability-emphasized-in-higher-ed/.

Hasnunidah, N., Susilo, H., Irawati, M. H., & Sutomo, H. (2015). Argument-Driven Inquiry with Scaffolding as the Development Strategies of Argumentation and Critical Thinking Skills of Students in Lampung, Indonesia. *American Journal of Educational Research*, *3*(9), 1185-1192

Iowa State University. (2016, June 1). Employers want college grads to have strong oral skills. *Science Daily*. Retrieved from https://www.sciencedaily.com/releases/2016/06/160601132144.htm.

Jackson, M. (1973). Debate: A neglected teaching tool. *Peabody Journal of Education,* 50 (2), 150-154.

Jagger, S. (2013). Affective learning and the classroom debate. *Innovations in Education and Teaching International*, *50*(1), 38-50.

Jerome, L., & Algarra, B. (2005). Debating debating: a reflection on the place of debate within secondary schools. *The curriculum journal*, *16*(4), 493-508.

Jones, E. A. (1995). *National Assessment of College Student Learning: Identifying College Graduates' Essential Skills in Writing, Speech and Listening, and Critical Thinking. Final Project Report*. US Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, *112*(3), 527-535.

Koklanaris, N., MacKenzie, A. P., Fino, M. E., Arslan, A. A., & Seubert, D. E. (2008). Debate preparation/participation: an active, effective learning tool. *Teaching and learning in medicine*, *20*(3), 235-238

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

Kwon, O. N., Bae, Y., & Oh, K. H. (2015). Design research on inquiry-based multivariable

calculus: focusing on students' argumentation and instructional design. *ZDM*, *47*(6), 997-1011.

Leite, C., Mouraz, A., Trindade, R., Martins Ferreira, J. M., Faustino, A., & Villate, J. E. (2011). A place for arguing in engineering education: a study on students' assessments. *European Journal of Engineering Education*, *36*(6), 607-616.

Leonhardt, D. (2014, May 27). Is College Worth It? Clearly, New Data Says. *The New York Times*. Retrieved from http://www.nytimes.com/2014/05/27/upshot/is-college-worth-it-clearly-new-data-say.html?_r=0.

Lilly, E. (2012). Assigned Positions for In-Class Debates Influence Student Opinions. *International Journal of Teaching and Learning in Higher Education*, *24*(1), 1-5.

Llano, S.M. (2015). Debate's Relationship to Critical Thinking. In M. Davies & R. Barnett (Eds.), *The Palgrave Handbook of Critical Thinking in Higher Education* (139-151). New York, NY: Palgrave Macmillan.

Nguyen, V. Q. C., & Hirsch, M. A. (2011). Use of a policy debate to teach residents about health care reform. *Journal of graduate medical education*, *3*(3), 376-378.

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, *328*(5977), 463-466.

Paris, B. (2016, November 29). Failing to Improve Critical Thinking. *Insider Higher Ed*. Retrieved from https://www.insidehighered.com/views/2016/11/29/roadblocks-better- critical-thinking-skills-are-embedded-college-experience-essay?mc_cid=e5fd55094b&mc_eid=9cde0a86f8.

Pitts, R. T., & Naumenko, O. (2016). The 2014 Standards for Educational and

Psychological Testing: What Teachers Initially Need to Know. *Working Papers in Education*, *2*(1), 1-6.

Proulx, G. (2004). Integrating scientific method & critical thinking in classroom debates on environmental issues. *The American Biology Teacher*, *66*(1), 26-33.

Rao, P. (2010). Debates as a pedagogical learning technique: empirical research with business students. *Multicultural Education & Technology Journal*, *4*(4), 234-250

Roy, A., & Macchiette, B. (2005). Debating the issues: A tool for augmenting critical thinking skills of marketing students. *Journal of Marketing Education, 27*(3), 264-276.

Scott, S. (2008). Perceptions of students' learning critical thinking through debate in a technology classroom: A case study. *The Journal of Technology Studies*, 34:1.

Stemler, S.E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation 9, 4*. Retrieved from http://pareonline.net/getvn.asp?v=9&n=4.

Tous, M. D., Tahriri, A., & Haghighi, S. (2015). The effect of instructing critical thinking through debate on the EFL learners' reading comprehension. *Journal of the Scholarship of Teaching and Learning*, *15*(4), 21-40.

van Rijn, P. W., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts. *Psicología Educativa*, *20*(2), 109-115.

Vo, H. X., & Morris, R. L. (2006). Debate as a tool in teaching economics: Rationale, technique, and some evidence. *Journal of Education for Business*, *81*(6), 315-320.

Winkler, C. (2011). To argue or to fight: Improving at-risk students' school conduct

through urban debate. *Controversia: An International Journal of Debate and Democratic Renewal*, (7:2), 76-90.

Yanklowtiz, R.S. (2013, August 15). A society with poor critical thinking skills: The case for 'argument' in education. *The Huffington Post*. Retrieved from http://www.huffingtonpost.com/rabbi-shmuly-yanklowitz/a-society-with-poor-criti_b_3754401.html.

Zarefsky, D. (2001). Argumentation. In R.O. Sloane (Eds.) *Encyclopedia of rhetoric* (Vol. 1). Oxford University Press on Demand.

Zarefsky, D. (2014). Rhetorical perspectives on argumentation. *Cham: Springer International Publishing*.