Youth Apprenticeship in Reasoned Discourse: The Power of Learning by Doing

Mariel Rebecca Halpern

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Youth Apprenticeship in Reasoned Discourse: The Power of Learning by Doing

Mariel Rebecca Halpern

Learning via apprenticeship is widely regarded as a powerful mechanism. To examine the role of apprenticeship learning and practice in developing argumentive thinking and writing, young adolescents (n = 64) participated in a four-week dialogic argumentation activity. They drew on available evidence and engaged 20 daily sessions in one-to-one electronic dialogues on contemporary social issues, anonymously, with a series of opposing-side partners. To assess the proposition that adolescents' argumentation skill advances via apprenticeship with a more skilled partner, in an experimental (but not control) discourse condition, a skilled adult arguer replaced a peer in half of the dialogues. Effects on students were evaluated in the dialogue and individual writing contexts. In the dialogue context, performance in initial peer dialogues during the first day of the workshop and in a final dialogic assessment on a new topic were evaluated. In the individual writing context, performance on the last workshop-debate-topic essay and non-workshop-debate topic essay were evaluated. Data were analyzed according to previously identified and well-validated coding schemes on counterargument and argument strategies. Although all participants showed skill gains, students in the experimental condition advanced in argumentive reasoning more rapidly than those in the peer-only control condition. Specifically, the strongest counterargument strategy (counter-undermine) appeared in greater proportions of idea units in the dialogues of students in the experimental condition, compared to those in the comparison condition. Only “weaken-other” improvements in dialogue reached significance in transferring to essays. These findings extend upon and support previous work on the power of

dialogic engagement and engagement with more competent others as a mechanism of apprenticeship learning. Pedagogical and social implications are discussed.

Table of Contents

List of Tables and Figures

Acknowledgements

I would like to acknowledge the extraordinary debt I owe to the individuals who have provided invaluable support towards my research and the gift of their time to help me grow over the years.

Dr. Deanna Kuhn, my academic advisor, you took me on as your doctoral student and gave me the freedom to explore an area of research that I'm genuinely interested in. My maturity and thinking about the development of these key 21st century skills have been augmented by the unparalleled professionalism and expertise that you bring to your research projects. As my thesis committee sponsor, you've been there every step of the way to support and guide me through the process, even as our lives were disrupted by the COVID-19 pandemic. I cannot thank you enough and aspire to conduct research with such skill, compassion, and precision as you have shown me.

I would like to express my deepest appreciation to my thesis committee. Dr. Bryan Keller, my committee chair and statistics professor, you've taught me to engage with data in ways that I never had before and helped me better understand how to approach thinking statistically on a conceptual level. Your input has and will continue to stimulate my intellectual curiosity and ambition. Dr. Christine Baron, I was delighted when you graciously agreed and made time to participate as a member of my committee. Your unwavering support through various refinements of my thesis has been invaluable to my research. Dr. Joan Lucariello, you've pushed me to reflect critically and conceptually on my theoretical framework. Dr. Anahid Modrek, you took the time to review my thesis with the utmost attention to detail and encouraged me to further consider the practical applications of my thesis. Thank you to all of

you. Your commentary and critiques have made my thesis richer and something that I am proud of having written.

I am also grateful to DK's lab members. Teresa Fraguada and Michael Youmans, I've thoroughly enjoyed "nerding out" with you, discussing course readings, sometimes in tremendous detail. Our talks have scaffolded my thinking. Sybille Bruun, thank you for your assistance with coding. I admire the candidly refreshing, warm, and enthusiastic energy you continually bring to the table. I must also thank Si Xiao and Yu Song for their superb effort and assistance with my thesis project. Flora Matos and Yuchen Shi, we first met at my accepted students day and you've shown me the ropes ever since. I am extremely grateful to all of you for our friendship inside and outside of the lab. You've offered intellectual and invaluable insight (and an escape, when needed) that I will take with me forever.

A special thanks to Dr. Irfan Ahmed Rind, who started his post-doc around the time I began my thesis work. Your dedication and commitment to scholarship, insight and intellect, and ceaseless compassion for others is truly inspirational. You're always willing to take time out of your busy schedule, even when facing a looming deadline, to lend an empathetic ear and offer perspective at challenging times. I've cherished our time together. It's a privilege to call you my academic older brother.

I cannot leave Teachers College without thanking Celia Goldsmith and the HUD office, as well as Russell Gulizia and Heidi Rizzo and the Office of Doctoral Studies for their continued administrative help and support. Their patience and understanding cannot be underestimated.

I hope I can pay forward what all of you have done for me.

M. R. H.

Dedication

This thesis is dedicated to my family.

To my parents, Jeannine and David, for your unconditional love and support, your incredible patience, and your endless dedication to your beliefs in my abilities.

To my brothers, Danny and Bobby, for your constant love and support and plain dependableness.

To Mike, for your selfless love, being my sounding board, and keeping me grounded.

To my grandparents, Yvonne and Patrick, for being my biggest cheerleaders.

The ways in which you all love and show up for those you love inspires me to be a better person.

I love all of you, with all of my heart.

Chapter 1: Introduction

According to data from the National Center for Education Statistics (National Assessment of Educational Progress [NAEP], 2011) Writing Report Card, only about 25% of middle and high school students wrote argumentative essays that effectively considered multiple viewpoints with supporting evidence. The PEW Research Center's Internet and American Life Project Online Survey of Teachers corroborate this finding: only about 15% of Advance Placement (AP) and National Writing Project (NWP) middle and high school teachers rated students abilities to synthesize materials and perspectives into a cohesive argument (Purcell et al., 2013). While most adolescents struggle with argumentative writing, teachers consider it critical for communicating responsibly and judging claims (Purcell et al., 2013)—competencies that are particularly relevant in today's connected world, where multidimensional, complex social structures not only complicate the kinds of knowledge and behaviors expected and demanded, but also call for different kinds of intellectual engagement that challenge traditional teaching and learning assumptions (Resnick et al., 2015).

Dialogic approaches to developing argumentative writing emphasize face-to-face and computer-mediated dialogic activities as a bridge to developing writing skill. Students engage in peer-to-peer discussion of ideas and issues of the past and present that afford opportunities to develop the cognitive and communicative competencies that underlie engaged citizenship (Kuhn et al., 2019; Mercer et al., 2020; Resnick et al., 2015). Empirical investigations of argumentation-based curricular interventions have led educational researchers to assert that

engaged, purposeful discourse on serious topics advances the argumentive[1] reasoning skill(s) that lie at the heart of critical thinking. Yet even with sustained practice, competencies develop slowly—middle school students only gradually begin to address alternative perspectives and coordinate claims and evidence (Asterhan, 2018; Kuhn & Crowell, 2011; Kuhn et al., 2013; Shi et al., 2019). Such gains are nonetheless robust and warrant further empirical investigation into the characteristics and conditions that promote and inhibit academically productive talk in various learning environments (Resnick et al., 2015).

Study Purpose and Research Questions

Existing work (Hemberger et al., 2017; Kuhn, 2018a; Kuhn, 2019; Kuhn et al., 2016a; Kuhn et al., 2019; Papathomas & Kuhn, 2017; Shi et al., 2019) suggests that the dialogic structure of the activity frequently makes its way into final essays students write on the topic (Kuhn & Modrek, 2021). The approach of having students talk directly to one another transfers a greater share of management of the discourse to students, relieving teachers of the burden of feeling that they must remain at the center of the conversation. Meanwhile, students gain an increasing sense of responsibility to one another and they come to embrace and uphold norms of discourse that this responsibility entails. The electronic mode allows students time to reflect on the accumulating exchanges that appear on the screen before them and plan their next move, promoting deeper discussion.

Rather than discussion of contemporary issues as "optional enrichment," discourse with peers about significant, challenging real-world issues should be an educational core and necessity, preparing students for futures that will depend on it. How else can they envision their

[1] *Argumentive* is a descriptive term (noun) used in earlier works by Kuhn and her colleagues to express disposition and function of argumentation, distinguished from *argumentative*, an action term (verb) used to express ability to perform argumentation. Both terms express a tendency toward the action of argumentation. Argumentive has become widely adopted in the literature in recent years.

future selves as informed, thoughtful contributors to debating solutions, especially today with as many poor public role models as good ones? Our youth must become involved as early as possible in contemplating the many issues their society faces. A potentially powerful mechanism, such as that of youth apprenticeship in reasoned discourse, is a first step to engage students in deep thinking and talking about the issues they might take action with regard to.

In sum, traditional curricula unconnected to students' personal, subjective realities leave students feeling disengaged from the real-world issues they will increasingly confront as they face participation as adult citizens in a democracy. In envisioning how to best help today's young minds understand deeply and feel empowered to address issues of their day, policy-makers and educational practitioners alike should embrace forms of learning that afford opportunities for adolescents to engage in reasoned debate and in so doing to develop the argumentation skills these activities require. The present work rests on the position that young people develop these skills by engaging in rich practice of them, with peers and especially, we propose, in experiencing and coming to adopt the skills of more able, experienced arguers. The study presented here serves to test this view.

To date, a few studies have investigated argumentation via apprenticeship learning, and those that have find that, with time, it serves as a powerful social and cognitive mechanism for advancing discourse skill (Mayweg-Paus et al., 2015; Papathomas & Kuhn, 2017; Zillmer & Kuhn, 2018). In the present study, we appropriate the term *apprenticeship learning*, to mean the dyadic, heterogenous (novice-expert) interaction between dialogic partners in the context of argumentation, and conceive of argument as a cultural activity mediated by language. Our appropriation of the term differs from the concept of activity in Soviet psychology that emphasizes practical (material) actions (e.g., activity is the basic unit of analysis used to

understand individual actions) (Kozulin, 1986) in that in-person cues are absent (all dialogic interaction is electronic in this experiment). It also differs from the role of cultural mediators of cognition in that the socially meaningful activities are wholly language mediated. For example, in *The Psychology of Literacy,* Scribner and Cole (1981) describe cultural practices to express the continuous ways in which children use their developing understandings of concepts introduced in natural settings (i.e., school or classrooms). In the present study, however, participants electronic (written) everyday talk (culture in specific) is the key source of problems to solve. Although the activities of apprenticeship learning take place outside of the formal classroom context, cultural forms and products, such as literacy or syllogisms, uniquely inform cultural processes of interaction and communication in the present study. Therefore, in this experiment, we investigate whether (1) an apprenticeship-based argumentation learning intervention develops argumentive reasoning and writing more effectively than engagement and practice alone with similar-skill peers, and (2) to what extent adolescents adopt skills they experience when interacting with more skilled arguers.

Chapter 2: Literature Review

Dialogic approaches draw on many theoretical traditions. It is thus critical to develop a clear conceptualization and operationalization of numerous key constructs. We attempt to clarify what is known about this web of interwoven theoretical constructs in this literature review chapter of theoretical and related empirical studies.

Argumentation as a Process

The term *argument* is used to reflect two kinds of argument: Argument as a product and argument as a process. To distinguish between the two kinds, we use the term *argument* to refer to argument as a product, an isolated act of rhetoric whereby an individual advances a claim using, minimally, one or more reason(s), and additionally, evidence and counterclaims, and the term *argumentation* to refer to argument as a dialogic process in which at least two individuals engage in constructing and evaluating competing rhetorical arguments with the goal to resolve competing claims (Iordanou & Rapanta, 2021; Kuhn & Udell, 2003; Kuhn et al., 2016a; Nussbaum, 2021). Put simply, a person makes an argument, but engages in argumentation; hence, argument is an individual act (intra-psychological), while argumentation is a social act (inter-psychological). Dialogic argumentation, therefore, is a means of cognitive engagement supported by theoretical conceptions that emphasize the complementarity of social and internal argument (Billig, 1987; Kuhn, 1991).

A Dialogic View

Dialogic approaches to education originate in some of the earliest theories of learning and teaching, going as far back as Socrates, to Vygotsky (1937/1987) whose theories of cognitive development set the roots for pedagogical practices oriented around the value of talk as we conceive it today, placing social interaction at the center of learning and development.

Philosophers, psychologists, linguists, and discourse specialists have all studied argumentation from a dialogic view and we briefly describe a few whose thinking has influenced how we frame the construct.

With respect to philosophical underpinnings, we draw on the work of Walton (2014), who referred to dialogue theory as "the underlying structure on which to base the analysis and evaluation of argumentation" (p. 1). From an argumentation theory point of view, the dual goals of argumentation, to increase the strength of one's own argument and weaken the force of the opposing argument, support progress in the procedural aspect (Walton, 1989). The fact that each dialogic partner has an influence on the other and partners' dialogic contributions overtime influence each other, underscores the need to evaluate arguments within their dialogic context, as partners seek to accomplish their goals through several strategies, e.g., counterarguments (van Eemeren & Grootendorst, 1992; Walton et al., 2008).

Within psychology, and specifically with respect to developmental origins, Piaget (1952) and Vygotsky (1937/1987) theorized that learning is an active process of "mental construction" whereby knowledge grows when new experiences, knowledge and knowing interact with that pre-existing (i.e., constructivism). In particular, Vygotsky's sociocultural framework focuses on the social and cultural constitution of learning and his core idea of collaborative cognition informs curricular approaches that emphasize using talk in effective ways for children's learning and development (Alexander, 2006, 2018; Asterhan & Schwarz, 2016; Kuhn et al., 2013; Mercer & Littleton, 2007; Mercer et al., 2020; Resnick et al., 2015; Reznitskaya & Wilkinson, 2017). More specifically, Vygotsky's belief that daily practice of discourse is a pathway of development for individual argumentive thinking, directly influences the dialogic view (Kuhn, 2019).

This theoretical view of thinking as a social practice has practical implications regarding how dialogic argumentation can be engaged in within educational settings in ways that will best support its development. One is that extensive practice and sustained engagement in argumentive discourse benefits thinking and learning. We explore this proposition and next, turn to an argumentation-based curricular intervention that supports development of argumentation skills with different cognitive and dialogic objectives.

A Dialogue-Based Pedagogical Approach: "Argue With Me"

"Argue With Me" (AWM) is a dialogue-based pedagogical approach developed by Kuhn and her colleagues (e.g., Kuhn et al., 2016a) that has garnered empirical support over the years for developing argument skills and dispositions in [primarily] adolescents (of very wide range and ability levels; see Iordanou and Rapanta (2021) for review). The AWM approach involves extensive practice in goal-based argumentation and reflection activities structured, sequentially, in three phases (Pregame, Game, Endgame), and is implemented by instructional coaches, twice weekly, over 13 class sessions, where students engage deeply with a series of challenging topics throughout the course of one or more school years (although several studies report intervention implementation in shorter durations, e.g., Iordanou et al. (2019)). Students typically work in same-side small groups during the AWM pregame, where they find reasons and evidences for their position. They engage in two forms of discourse during the game, first verbal, where same-side peer-pairs prepare for electronic dialogues with opposing-side peer-pairs. This second form of electronic discourse provides a written record of externalized thought, facilitating reflection on what is exchanged. Throughout the game, instructional coaches share pieces of information in question-and-answer form (students are also encouraged to ask questions of their own). During the endgame, students work in same-side large groups as they prepare for the whole class

"showdown" debate. An individual essay assignment serves as the intervention's final culminating activity. Within the past decade, a substantial body of research evidence has documented the effectiveness of the AWM approach amongst adolescence.

AWM promotes gains in argument skill, such as counterargument (Crowell & Kuhn, 2014; Kuhn et al., 2008; Kuhn et al., 2016b; Kuhn et al., 2019; Mayweg-Paus et al., 2015; Papathomas & Kuhn, 2017) and rebuttals (Iordanou, 2013); epistemological gains, such as epistemic understanding (Kuhn et al., 2013) and intellectual values (Kuhn et al., 2011); meta-cognitive gains, such as meta-talk (Kuhn & Zillmer, 2015; Zillmer & Kuhn, 2018); and gains in use of information as evidence (Shi, 2019), such as coordinating claims with evidence (Hemberger et al., 2017; Kuhn & Moore, 2015). More recently, Iordanou et al. (2019) dissected the multicomponent intervention to isolate the powerful question-and-answer method and test knowledge gains. Iordanou and Kuhn (2020) isolated dyadic interaction with opposing-side peer-pairs in comparison to same-side peer-pairs and found the former beneficial. Most empirical investigations of AWM also test and observe transfer, typically either across contexts (e.g., dialogue to individual essays) or content/topic knowledge. Nonetheless, the benefits of the AWM approach span various aspects of cognitive development and provide a comprehensive body of evidence for AWM's conceptual foundation that the dialogic process is a path to individual written argument (Hemberger et al., 2017; Kuhn & Crowell, 2011; Kuhn et al., 2016a, 2016b; Shi et al., 2019).

Dialogue as a Path to Individual Written Argument

The AWM approach is founded on the key principle that dialogue serves as a bridge to individual argumentive writing (Hemberger et al., 2017; Kuhn & Crowell, 2011; Kuhn & Modrek, 2021; Kuhn et al., 2016a, 2016b; Matos, 2021; Shi et al., 2019). The strength of the

approach lies in exploiting the dialogic exchanges that come naturally to children by extending their engagement in it. Sustained practice provides purpose: it allows two parties to speak or write directly to one another, providing the otherwise "missing interlocuter" (Graff, 2003), along with a shared responsibility for maintaining the exchange (Shi et al., 2019). Its power lies in the fact that it effectively removes the teacher, who traditionally directs classroom discourse by posing questions and identifying speakers, from the students' interaction, which transfers opportunity for students to figure out what their message is, not what their teacher wants to hear (i.e., reasoning rather than rote responding). Most notably, the structural features of the dialogic context shift traditional classroom cultural scripts and discourses that shape students' understandings of what "counts" as valuable, yielding appreciation of intellectual values. To summarize this fundamental concept, dialogue as a path to argumentive writing provides a clearly defined audience on the one hand, and meaningful purpose on the other—two components critical to successful writing (Shi et al., 2019).

What evidence do we have that dialogue serves as a path to developing expository writing? To date, 29 empirical studies with students of different backgrounds and across different spans of development (childhood, adolescence, and adulthood) used the AWM approach to investigate this question (Iordanou & Rapanta, 2021). Two main lines of evidence have emerged, one concerns changes in novice arguers' essays on successive topics over the course of one or more years (Hemberger et al., 2017; Kuhn et al., 2016a; Shi et al., 2019) and the other direct comparison of students' dialogues and essays (Iordanou, 2013; Kuhn & Moore, 2015; Papathomas & Kuhn, 2017).

Studies show that young adolescents not only increasingly use evidence in their writing tasks, but also increasingly use different kinds of evidenced-based claims overtime and with

extended opportunity to practice (Hemberger et al., 2017; Kuhn et al., 2016a, 2016b; Shi et al., 2019). At the end of one year of participation in AWM, novice arguers use information as evidence consistently in their final essays that typically function in support of a claim (Hemberger et al., 2017; Kuhn & Moore, 2015; Kuhn et al., 2016a; Shi et al., 2019). Only later and with continued practice do students steadily draw on another type of evidence that functions to weaken the opposing claim and at even later stages, integrate evidence typically used in support of the opposing position and rarely, used to weaken their own position (Crowell & Kuhn, 2014; Hemberger et al., 2017; Shi et al., 2019).

Research has documented the dialogic structure of students' argumentation with peers making its way into their essays, but progress manifests in dialogues before it does in writing individual essays (Hemberger et al., 2017; Kuhn & Moore, 2015; Kuhn et al., 2016a, 2016b; Papathomas & Kuhn, 2017; Shi et al., 2019). For example, Hemberger et al. (2017) and Shi et al. (2019) manipulated the sequence of pieces of Q&A information, sharing first, "support-own", then "weaken-other", followed by "support-other", and "weaken-own", and found that when instructional coaches shared "weaken-other" Q&A information, novice arguers typically used it frequently in dialogue, but only later in their topic essays (less than 15% used "weaken-other" at least once in the first topic essay and more than 70% in the final essay). Additional investigations comparing dialogues and essays reveals that novice arguers tend to draw on prior personal information as evidence in dialogues, but shared Q&A information in essays (Kuhn & Moore, 2015; Macagno, 2016).

Although there is still much to learn about the conditions in which developmental progress is best supported, research evidence supports development from the social context of dialogues to the individual context of essay writing. In addition, findings support developmental

progress in the direction of integrative complexity—from *single-* to *dual-* focus essays in which evidence is used in service of claims that support one's own position and weaken the opposing position, to *integrative-* focus essays exemplified by the "however" structure that connects supporting the opposing position and/or weaking one's own position with the aforementioned two types ("support-own", "weaken-other")—that reflects meta-level understanding (Hemberger et al., 2017; Kuhn & Moore, 2015; Kuhn et al., 2016a, 2016b; Papathomas & Kuhn, 2017; Shi et al., 2019) and reinforces the power of the dialogic perspective.

Mechanism of Development

Dialogic approaches originate in Vygotskian and Piagetian theories of cognitive development which set the roots for contemporary educational psychology and directly connect to educational research on argumentation skills. According to the Vygotskian sociocultural perspective, regulation is social in nature and internalized or interiorized (Kuhn, 2018a; Kuhn, 2019; Kuhn et al., 2016b) to become an individual process. The more capable partner provides support for the less capable partner, creating a *zone of proximal development* (i.e., the distance between the less capable partners actual abilities and potentially achievable abilities) (Vygotsky, 1978). According to the Piagetian sociocognitive perspective, regulation is an individual process influenced by contextual and social aspects. Dialogic partners of similar abilities interchangeably take on the role of providing and receiving support as needed, flexibly scaffolding learning (Iordanou & Rapanta, 2021).

Recent research using the AWM approach shows that dialogic engagement with more capable others (a condition emphasized in the Vygotskian framework), as well as with peers of similar ability (a condition emphasized in the Piagetian framework) enhances argumentation skill (Mayweg-Paus et al., 2015; Papathomas & Kuhn, 2017). Moreover, Mayweg-Paus et al. (2015)

and Papathomas and Kuhn (2017) found cognitive gains in the quality of argumentation occurred significantly sooner and quicker when novice arguers participated in argumentive discourse modeled by an adult "expert" arguer (i.e., a trained, qualified member of the research team) and concluded that apprenticeship learning is a powerful developmental mechanism for advancing discourse skill (although the researchers did not tell students that they substituted the opposing-side peer-pair during the game phase e-dialogues). Still, the fact that the benefits of peer collaboration extend to equal- and unequal- ability peers, suggests that both sociocognitive and sociocultural processes have a place in dialogic approaches to teaching and learning.

Scholars studying argumentation skill development besides the AWM approach report benefits of dialogic approaches to teaching and learning through neo-Piagetian developmental mechanisms. For example, Nussbaum and his colleagues have used Argumentation Vee Diagrams (AVD), graphic organizers that scaffold integrative reasoning and specifically support *argument-counterargument integration*, the process of evaluating both sides of an issue to reach an overall final conclusion (Nussbaum, 2008; Nussbaum & Schraw, 2007). AVDs encourage students to consider criteria for weighing arguments through Critical Questions (CQs) to organize writing of reflective opinion essays (Nussbaum, 2021; Nussbaum & Putney, 2020; Nussbaum et al., 2019). Students list arguments and counterarguments on a central question on the respective sides of the vee, then use CQs to evaluate the opposing arguments they have listed (Nussbaum et al., 2019). In a similar vein, Lombardi and his colleagues have used Model-Evidence-Link (MEL) diagrams developed by Chinn and Buckland (2012) in which students, working together in small groups, draw arrows to indicate whether specific pieces of evidence support, contradict, or have nothing to do with, two competing models on socioscientific issues

(e.g., climate change). Lombardi et al. (2013, 2016) found that MEL diagrams act as scaffolds for students to revise plausible perceptions and reasoning.

Contemporary education theorists focus their research on the educational power of discourse through a Vygotskian lens. Resnick and Mercer and their respective colleagues (Littleton & Mercer, 2013; Mercer & Howe, 2012; Resnick et al., 2015; Resnick et al., 2018) emphasize the value of discourse engagement as a practice in its own right. For example, Resnick et al. (2015) describe *accountable talk*™ (Michaels et al., 2008; Resnick et al., 2015; Resnick et al., 2018), a phenomenon that reflects the idea that deep learning is accountable to reasoning, knowledge, and the learning community. Mercer (1995) identifies *exploratory talk* as a type of talk in which partners engage critically and constructively with each other's ideas within the norms of the discourse. Alexander (2006, 2018) describes a dialogic teaching framework focused on the teacher-student interactions to promote classroom dialogue. Along similar lines, Reznitskaya and her colleagues advocate using a specific technique focused on the role of dialogue in literacy instruction with elementary grades classrooms. Reznitskaya et al. (2009) and Reznitskaya and Wilkinson (2017) describe *inquiry dialogue* as a technique where teachers encourage their students to search for the most reasonable response to a contestable question through collaborative argumentation. During discussions of assigned readings, teachers draw upon a repertoire of *talk moves* that they use to help their students take positions on issues, support positions with reasons and evidence, and challenge others positions, enhancing students' dialogic experiences albeit minimal transfer (Reznitskaya et al., 2012). Somewhat similarly, Murphy et al. (2018) describes teacher-facilitated *quality talk* in fourth-grade classrooms to promote comprehension.

Differences in development may be explained by the differences in approaches. The mode of dialogic interaction in AWM is both verbal and written, whereas the mode is mostly verbal when using AVDs and MEL diagrams. Additionally, whereas dialogic interaction typically occurs in small group work and most empirical investigations of AVDs and MELS are conducted with undergraduate or high school students, respectively, interaction occurs in peer-pairs with mostly middle school students in AWM (and in the online adaptation described in the present study, young adolescents engage in dialogue, one-to-one with an opposing-side partner). Similarly, dialogic teaching and learning approaches that emphasize talk in the classroom often have students working in groups (whole class, large group, and/or small group) and focus on the student-teacher dialogic interactions that take place. While most empirical investigations occur with adolescents, the above-mentioned literacy studies took place with elementary-aged students. Distinctions in these approaches likely influence different dialogic experiences and learning outcomes and development.

Regardless, in both traditions (sociocognitive or sociocultural) the dialogic interactions that take place (i.e., *social construction*) and the individuals learning that is influenced by the dialogic context (i.e., *social influence*) have a role to play and demonstrate the complex nature of psychosocial phenomena. Constructivist processes of coordinating schemas into integrated structures and internalization, among other things, support developmental progress in integrating complex arguments and advancing argumentation skill. Nevertheless, there is much to learn regarding the mechanisms through which advancements in argumentation best occur.

Assessing Argument Skill

Individual argument skill presumably develops from dialogic interaction, however, development of argumentive discourse is difficult to study as an individual skill because the unit

of analysis is not the individual. As such, tracing the evolution of argumentive writing in both dialogic and individual contexts is complex, which makes addressing the relationship between peer argumentation and individual argumentation and reasoning skills statistically challenging. As a result, many dialogic pedagogy researchers analyze data by developing coding schemes, training coders, coding, and analyzing, among other things, that provide insight into cognitive development through qualitative and quantitative indices.

Previous empirical work using AWM (e.g., Crowell & Kuhn, 2014; Mayweg-Paus et al., 2015; Papathomas & Kuhn, 2017, etc.) has used different argument strategies to shed light on the underlying structure of argumentation. In these empirical investigations, the author(s) initially devised a coding scheme that draws on Walton's (2005) coding scheme that emphasizes the proportion of utterances that weaken the opposing side's position to detect and classify the types of argumentation strategies observed (i.e., counterargument strategies). In addition to this emphasis on counterargument, other studies (e.g., Felton, 2004; Felton & Kuhn, 2001; Macagno et al., 2014; Zillmer & Kuhn, 2018) have assessed development on the increasing use of more sophisticated discourse moves (e.g., meta-talk, questioning) overtime and with sustained engagement. Several studies have also focused on the functional use of evidence (e.g., Iordanou et al., 2019). Each of these aspects of development are important for development of discourse skill and are relevantly measured in the present study. In this dialogic framework, the present study investigates the hypothesis that novice participants who interact with an expert arguer benefit the development of more sophisticated argumentation strategies.

The Present Study

The present study extends upon Mayweg-Paus et al. (2015) and Papathomas and Kuhn (2017) investigations of dialogic interaction with a more capable other. Although the study

conducted by Mayweg-Paus et al. (2015) was more limited and conducted with a different sample and population of sixth grade students than Papathomas and Kuhn (2017), both studies compared peer-only dialogic interaction with expert-peer dialogic interaction by evaluating students' counterargument strategy use. Unlike the present study, both Mayweg-Paus et al. (2015) and Papathomas and Kuhn (2017) occurred in classroom contexts implementing the AWM approach. The present study was conducted in an all-online environment presented to participants as an extracurricular summer activity. Additionally, following the full, in-person AWM approach, the students in Mayweg-Paus et al.'s (2015) and Papathomas and Kuhn's (2017) studies first engaged in verbal discourse, working in same-side peer-pairs to agree on what they wanted to say in text, through e-dialogues (e.g., google chat), to their opposing-side peer-pair. This means that in both of the earlier studies, students engaged in both forms of discourse—verbal, same-side and written, opposing-side—but in the present study, all discourse occurred electronically, in writing and accordingly, dialogic interaction was only one-to-one, with the partner holding the opposing position. Because we were interested in individual participants' skill development and because dialogic interaction was one-to-one, we analyzed each individual participant's performance and development across dialogic partners and time (this means that we treated partners in the dyad as separate individuals, and analyzed only one individual per dialogue) for theoretical and analytical purposes.

Chapter 3: Method

Participants

Participants were 64 rising 6th through 9th graders. Of the 64, 15.6% (n = 10) were in the 6th grade, 26.6% (n = 17) were in the 7th grade, 21.9% (n = 14) were in the 8th grade, and 12.5% (n = 8) were in the 9th grade (data were missing for 23.4% (n = 15)). Males made up more of the sample than females: 55% (n = 35) were males and 45% (n = 29) were females. Zero participants self-identified as nonbinary. Participants attended public and private schools located in mostly urban, but some suburban areas of the Northeast United States. Data on Socio-Economic Status (SES) and Racial-Ethic distribution were not collected. However, we used data from the National Center for Education Statistics (NCES) Common Core of Data (CCD) for the 2019 to 2020 school year that collects relevant data annually from state education agencies, along with demographic data from private schools and the National Association of Independent Schools (NAIS) to determine demographic information of the schools represented in the sample, as an approximate. We were able to collect and examine data for 86% of the 22 total schools represented by the 64 students in the sample (data were missing for 13 participants). Genders were roughly equally represented in the school populations, with slightly higher percentages of males (52%) and slightly fewer percentages of females (48%). Of the student bodies, approximately 39% were White, 13% Black, 23% Hispanic, and 41% qualified for free or reduced-price lunch.

Participants learned about the workshop via a flyer that was distributed to principals and parent coordinators of urban public schools, located in the Northeast United States. Interested families registered for the workshop by mailing a note, signed by the middle school participant(s) and their parent/guardian, agreeing to attend at least four of five weekday sessions

conducted virtually (i.e., 16 or more sessions over four weeks), along with a $20.00 registration fee to secure their spot. Participants who completed 16 sessions received a certificate of commendation from the Teachers College Education for Thinking Center and those who completed 20 sessions also received a $25.00 Amazon gift card.

Procedure

Participants attended a four-week Debate-and-Decision-Making Workshop during the summer and fall of 2020. The workshop, offered by the Education for Thinking Center at Teachers College, is an online adaptation of the dialogic argumentation-based curriculum described in *Building Our Best Future: Thinking Critically About Ourselves and Our World* (Kuhn, 2018b), which features the AWM approach and activities designed specifically for adolescents.

Prior to the beginning of the workshop, participants were asked to submit a brief essay indicating which side they preferred and why for the initial topic, making it possible, most of the time, to assign a student to debate the topic with another student who held the contrasting view. Participants were also assigned a unique animal name by which they would be identified throughout the workshop.

During each of the four weeks of the workshop, two half-hour sessions took place at a fixed hour-long period each day. At the initial session, participants were introduced to the coordinator and her assistant, the activities that would take place, and procedures for interacting with the software. The dialogues then began during the second half-hour. These continued in a regular way, with new partners assigned for each dialogue, over the four-week period. At the end of the session, participants completed an evaluation of the dialogue where they were asked to rate their partners performance (“how good an arguer was your partner today?”) on a scale of

zero to 100. The purpose of this activity was to test if the participants perceived the manipulation. Beginning with the third session of each week (second day), four to six pieces of information presented in a question-and-answer format were made available, a few each day. These were accessed by students if and when they wished from a website they were instructed how to access. The purpose of these, students were told, was to get information that might be useful in making their arguments. Students were further told, "While waiting for your partner to respond will be a great time to see answers to these questions or get answers to any other questions of your own you may have." At the final session of each week, the coordinator offered some general group-level feedback, and, as a culminating activity, participants individually wrote and submitted a final position piece on the topic.

Each week, participants addressed a new topic. These were, in the order encountered, teen justice system (should teens who get in trouble with the law be dealt with in adult court or a juvenile court for teens?), college versus work (when you finish high school, should you have the choice of going to college or working for a few years first?), animal research (should animals be used in research to test new drugs, medical procedures, or other devices?), and effort allocation (should you put most time and effort in being at the top of your class in your good subject or in getting better in your poor subject?).

Posttest Assessment

During the week following the end of the workshop, participants were asked to complete two additional tasks scheduled at their convenience. Both related to a fifth topic, whether undocumented immigrants should be forcibly removed or allowed to stay, that had not been part of the workshop. The purpose of both was to assess any generalization of skills practiced in the workshop to new content. One was an essay on the topic, which participants were asked to

submit (in lieu of the final week's topic essay); the other was to participate in an electronic dialogue on the topic with an anonymous partner, who was a member of the research team. Their contribution to the dialogue needed to be specially constructed so as to provide a consistent input across participants and thereby minimize variability, allowing an assessment of the participant's dialogic skills with minimal variation across participants in the interlocutor's influence on participants' dialogic performance. In these dialogues, the template the adult interlocutor followed was to offer a standard sequence of reasons for favoring the position opposing that of the participant. Initial dialogues were all randomly assigned peers. A template for the initial dialogue was not considered necessary as all participants were equivalent in lacking any previous experience and variation within or across conditions was minimal (see means and standard deviations reported in Results section); nor was doing so feasible as initial dialogues took place simultaneously.

Intervention Manipulation

One third of participants (n = 26, 10 female, of the final sample of 64), chosen randomly, served in an experimental group and the remainder in a comparison group. Participants in the experimental group, unbeknownst to them, in roughly half of their intervention dialogues interacted with an adult who was a rotating member of the research team, referred to here as experts. The remaining dialogues occurred with rotating peers, as did all dialogues of those in the comparison group. The adult expert followed general guidelines but their contributions were not scripted (as they were in the posttest dialogue). While also identifying their own positions, opposing that of the participant, and offering suitable supporting arguments when the participant asked them to, the expert focused on exploring and later countering the participant's assertions, requesting justifications and empirical support of these with frequent use of "How do we know?"

as a query. Counters questioned both empirical correctness (“Do we know that’s true?”) and the participant’s reasoning (“If this is true, does it necessarily follow that. . .”).

Of the 64 participants, 57 (89%) completed an initial dialogue and 47 (73%) completed a posttest dialogue. (The 11% not completing an initial dialogue did not participate in the initial day of the workshop due to forgetting to attend or to technical difficulties signing in; the 27% not completing the posttest dialogue were unavailable the week following the workshop.).

Chapter 4: Results

The results chapter comprises three sections. The first section presents results of participants' ratings of the dialogues for the primary purpose of a manipulation check. The second section presents the results of an analysis of students' initial and final dialogues to assess change over time and effects of condition. The final section is addressed to students' individual essays.

Dialogue Ratings by Participants

As a manipulation check, a dependent samples *t* test was conducted for participants in the experimental condition on the difference scores between the average rating participants assigned to their expert-partners and the average rating assigned to their peer-partners. The specific research question that we set out to answer with this analysis was whether the participants in the experimental condition perceived the manipulation, as indicated by higher average total ratings of expert- dialogic partners than peer- dialogic partners, on a scale from zero to 100. The majority of the participants ratings of peer-partners fell within the range of 16 to 93 and ratings of expert-partners fell within the range of 30 to 97, on the zero-to-100-point scale. The mean value of participants' ratings of their expert-partners was 74.68 (SD = 18.89) and ratings of their peer-partners was 65.53 (SD = 21.51), for a mean difference of 9.15 (SD = 15.52), a statistically significant difference, $t(19) = 18.10$, $p < .001$. Therefore, the null hypothesis was rejected in favor of the research hypothesis that participants' recognized the superior quality of an expert's, compared to a peer's contributions to a dialogue.

As an additional analysis, we conducted an independent samples *t* test on participants (in both conditions) ratings of peer-partners. The independent variable was condition. On average, ratings of peer-partners ranged from 21 to 96 for participants in the comparison condition and 16

to 93 in the experimental condition, on the zero to 100 point scale. The mean value of participants' ratings of their peer-partners was 64.90 (*SD* = 19.33) for participants in the comparison condition and 65.53 (*SD* = 21.51) for the participants in the experimental condition, for a mean difference of 2.87. This difference was not statistically significant, $t(52) = 0.39$, $p = 0.27$. The participants in the experimental condition did not rate their peer partners differently than those in the comparison condition.

Initial and Posttest Dialogue Analysis

The dialogues chosen for comparison across time and condition were the initial peer dialogue the participant engaged in and the final (posttest) dialogue. These were chosen so as to equate as closely as possible the two dialogues to be examined. The two selected are the only two dialogues that participants had not previously discussed (at least within the confines of the workshop) and had not yet been presented any information with regard to that could potentially be used as evidence. The intervention manipulation began after the initial peer dialogue and ended before the final dialogue (i.e., only initial ("pre-") and final ("post-") assessment data were analyzed).

The unit of analysis remained the utterances of a single participant in the dialogue, rather than the pair, and only what this participant said was the subject of analysis, given the research hypothesis addresses the skill development of an individual. As noted earlier, in the posttest dialogue the interlocutor was an adult following a prescribed template, and in the initial dialogue the partner was a randomly chosen peer but individual variation was slight, given all participants were novices at this point with regard to the activity and the topic.

The statistical analysis consists of a mixed two-factor analysis with one between-subjects two-level factor (Experimental and Comparison) and one within-subjects two-level factor (Initial

and Final Dialogue). A planned pairwise comparison was conducted for the between-subjects factor during the posttest dialogue (i.e., the second level of the within-subjects factor). Post-hoc pairwise comparisons were conducted where the interaction was significant and for the within-subjects factor within each level of the between-subjects factor.

An additional analysis of proportion of use, i.e., the number of participants who used an argument strategy in their dialogues at least once, was conducted to assess if the extent of use of an argument strategy differed between groups. Pearson's Chi-square Test(s) of Independence were conducted to explore these relationships. A Fisher's Exact Test of Independence was used in cases where the Pearson's Chi-square Test of Independence could not be used due to low expected counts in two-by-two tables. For the purposes of failing to reject or rejecting the two-tailed null hypothesis, the significance level was set at an alpha level of 0.05.

Identifying and Coding of Idea Units

All participant identifying information was removed prior to coding to preserve privacy and minimize bias. Dialogues were coded by the author and a research assistant who was familiar with the study but blind to experimental condition. All contributions to the dialogue made by the participant whose utterances were being coded were first segmented into idea units, i.e., utterances that convey a single idea. Acceptable inter-rater agreement of 93% was achieved among two raters on 25% of the data, Cohen's kappa = 0.85, $p < .001$.

Idea units were coded into categories that identify the functional relation of the idea unit to the immediately preceding utterance of the dialogic partner. The specific functions identified in Table 1, can be divided into three broader groups, *Probing*, *Substantiating*, and *Countering*, plus an advanced category, *Concession*, that acknowledges some merit in the opponent's claim or weakness in one's own claim. The coding scheme is adapted from one described by Macagno

(2016) and by Papathomas and Kuhn (2017) and in earlier versions by others (Crowell & Kuhn, 2014; Mayweg-Paus et al., 2015). Percentage agreement between the same two coders was 91%, Cohen's kappa was 0.86, $p < .001$. Disagreements were resolved through discussion. The remaining dialogues were coded by the author.

The two functions in the Probing category, *Question* and *Meta-talk*, play a fundamental role in argumentation in that they can shift it to a deeper level, preparing the way for counterargument. Questioning seeks explicitly to find out more about the other's view. Meta-talk addresses and reflects on the discourse itself. It can take the form of a question or a statement.

Substantiating an assertion with evidence also plays a fundamental role in serving to support (or weaken) claims. Information is not automatically evidence. The arguer must identify when and how information has the potential to function as evidence and then coordinate it with an appropriate claim.

Countering constitutes the core of argument, if successful weakening the force of an opponent's argument. *Counter-Disagree* and *Counter-Alternative* are the weaker two of the four forms of countering because they leave the opponent's claim unaddressed. Counter-disagree, the weakest form, simply expresses disagreement without justification (e.g., "I disagree"). Counter-alternative goes beyond simple disagreement by advancing a different argument, one that leaves the opponent's claim unaddressed, hence failing to weaken it.

Counter-Critique seeks to weaken the opponent's argument by critiquing the opponent's preceding claim as incorrect. *Counter-Undermine*, the strongest form of counterargument, seeks to weaken the opponent's argument by undermining the opponent's reasoning, specifically the link between premises and conclusions.

Concession neither counters the opponent's argument nor concedes it is correct, but rather concedes that it has some merit, typically in the broader context of countering it, or concedes that one's own position has some weakness, despite endorsing it.

Examples of the specific functions these idea units serve appear in Table 1.

Table 1

Summary and Examples of Functional Types of Idea Units in Analytic Scheme for Coding Utterances in Argumentive Dialogues

Category	Description	Examples
Probing		
Question	An utterance that requests a response from the dialogic partner (*Q*)	"Could you elaborate?" "What are your reasons?"
Meta-talk	An idea regarding the dialogue itself (*Meta*)	"What does that have to do with our topic?" "I understand that is your opinion but this is a debate, and we are supposed to argue against the other person's opinion." "We never know because we don't have statistics."
Substantiating		
Evidence	A factual statement intended to strengthen or weaken a claim (*E*)	"There is room for another 10 million people." *(in response to, "There's not enough room in the US for everyone who wants to come here.")* "The population has been declining for 20+ years." *(in response to,"There's not enough room in the US for everyone who wants to come here.")*
Countering		
Counter-Disagree	A form of counterargument that rejects the opponent's argument without providing a justification for doing so (*Counter-D*)	"I don't agree with you."

Category	Description	Examples
Counter-Alternative	A form of counterargument that contradicts the opponent's argument by introducing an alternative argument (*Counter-A*)	"They benefit from being here so it's helpful for them to be here to do work and things like that" *(in response to, "There's not enough room in the US for everyone who wants to come here.")*
Counter-Critique	A form of counterargument that critiques the opponent's preceding claim as incorrect (*Counter-C*)	"The USA is a pretty large country and there is a lot of space for a lot of people" *(in response to, "There's not enough room in the US for everyone who wants to come here.")*
Counter-Undermine	A form of counterargument that undermines the opponent's reasoning (*Counter-U*)	"It won't discourage others because their needs are too great*" (in response to, "Sending them back will discourage others from coming.")* "They came here for a better life, so they wouldn't be better off back home" *(in response to, "They'd be better off back in their own home country.")*
Concession	Acknowledgement that the opponent's claim has some merit or one's own claim some weakness	"For some people maybe…" *(in response to, "They'd be better off back in their own home country.")* "It sounds mean [to send them back] but we need to keep everyone safe" *(in response to stay argument).*

Note. Examples come from posttest dialogues in which participants addressed the topic "Should young people brought illegally to the US as children be allowed to stay or sent back?"

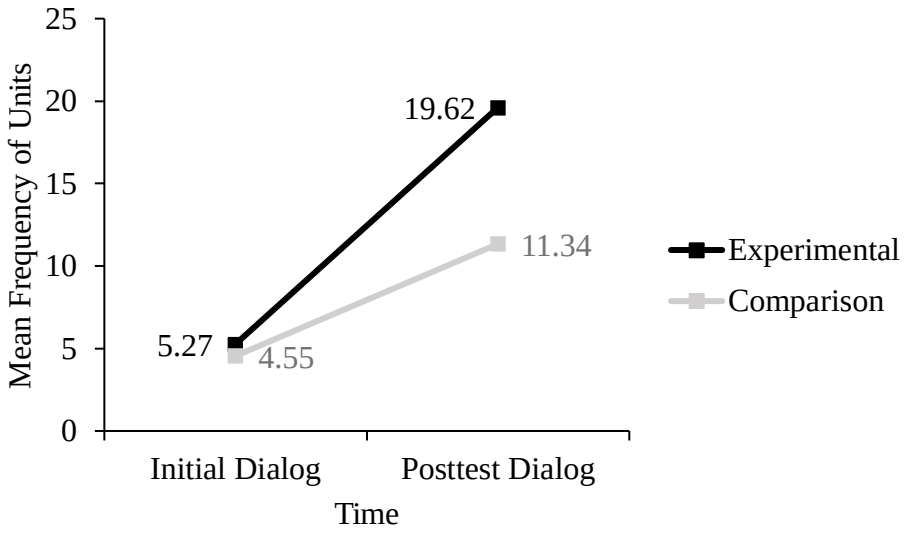
Initial and Posttest Dialogue Performance by Condition

Idea Units. Mean number of idea units contained in initial dialogues were similar across conditions: $M_{Experimental}$ = 5.27 ($SD_{Experimental}$ = 3.16); $M_{Comparison}$ = 4.55 ($SD_{Comparison}$ = 3.41). Means were higher at posttest in both conditions, reflecting a significant effect of time, $F(1, 62)$

= 49.74, $p < .001$, but diverged by condition, $F(1, 62) = 7.88$, $p = 0.01$, yielding a significant interaction between time and condition, $F(1, 62) = 6.36$, $p = 0.01$ (see Figure 1).

Figure 1

Mean Number of Idea Units by Time and Condition



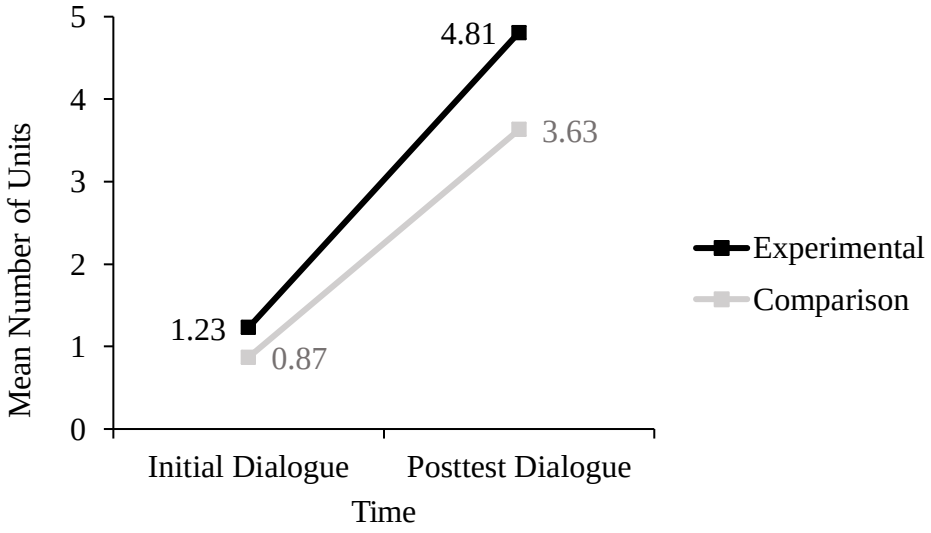
A planned pairwise comparison using the Bonferroni corrected alpha level of 0.025 revealed that participants in the experimental condition had a significantly higher mean number of idea units during the posttest dialogue compared to the participants in the comparison condition, $t(25.83) = 2.79$, $p = 0.01$ (Note: Levene's Test for Equality of Variances was violated for $F_{Posttest\ Dialogue} = 14.52$, $p < .001$). Post-hoc pairwise comparisons revealed that participants in the experimental condition had significantly more mean number of idea units during the posttest dialogue compared to the initial dialogue, $t(25) = 5.28$, $p < .001$. There was also a significant difference between the mean frequency of use of the number of idea units in the comparison condition, $t(37) = 4.14$, $p < .001$.

Probing. Initial and posttest dialogues were examined for the two probing subtypes, questions and meta-talk. In the experimental condition, the mean frequency of use of questions

during the initial dialogue was 1.23 (*SD* = 1.68), which increased to a mean of 4.81 (*SD* = 4.63) during the posttest dialogue. In the comparison condition, mean frequency of use during the initial dialogue was 0.87 (*SD* = 1.53) which increased to a mean of 3.63 (*SD* = 3.96) during the posttest dialogue. There was a significant effect of time, $F(1, 62) = 34.50$, $p < .001$, but a non-significant effect of condition, $F(1, 62) = 1.58$, $p = 0.21$ and non-significant interaction, $F(1, 62) = 0.57$, $p = 0.45$ (see Figure 2).

Figure 2

Question Subtype Usage by Time and Condition



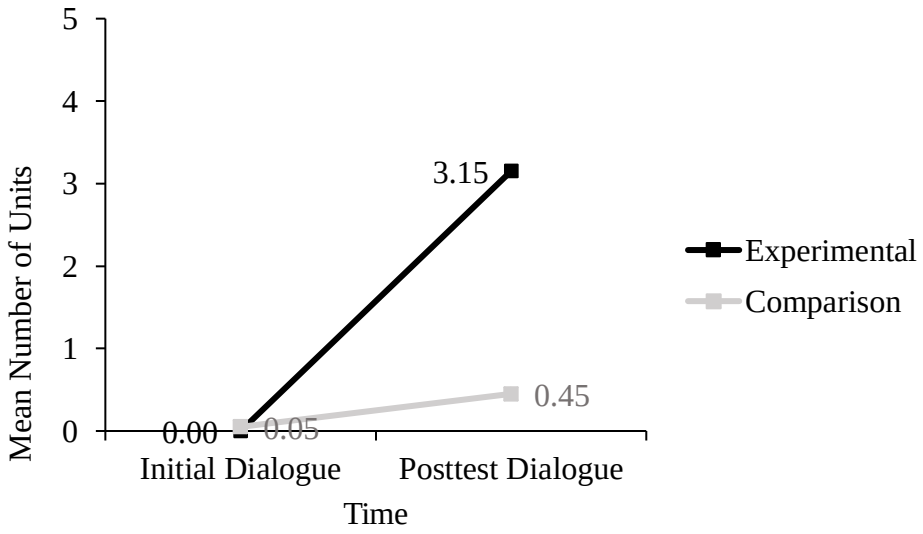
A planned pairwise comparison using the Bonferroni corrected alpha level of 0.025 was non-significant, $t(45) = 0.42$, $p = 0.68$. Post-hoc pairwise comparisons revealed that participants in the experimental condition used questions significantly more during the posttest dialogue compared to the initial dialogue, $t(25) = 4.24$, $p < .001$. There was also a significant difference between the mean frequency of use of questions in the comparison condition, $t(37) = 4.06$, $p < .001$.

The proportion of participants in the experimental condition who displayed the question subtype at least once was 52% (n = 13) at initial time and 56.3% (n = 18) at final time, compared to 95.2% (n = 20) at initial time and 96.2% (n = 25) at final time in the comparison condition. These differences were non-significant, $X^2_{Initial\ Dialogue}(1, N = 57) = 0.10$, $p = 0.75$; $X^2_{Posttest\ Dialogue}(1, N = 47) = 0.02$, $p = 1.00$.

Participants in the experimental condition used meta-talk during the posttest dialogue on average 3.15 (*SD* = 5.73) times, an increase from 0.00 (*SD* = 0.28), during the initial dialogue. The comparison group used meta-talk during the posttest dialogue on average, 0.45 (*SD* = 1.59) times, compared to 0.05 (*SD* = 0.73) times during the initial dialogue. The interaction between time and condition was significant, $F(1, 62) = 7.75$, $p < 0.01$, as was the effect of time, $F(1, 62) = 12.82$, $p < .001$ and of condition, $F(1, 62) = 7.23$, $p = 0.01$ (see Figure 3).

Figure 3

Meta-talk Subtype Usage by Time and Condition



A planned pairwise comparisons using the Bonferroni corrected alpha level of 0.025 revealed that the participants in the experimental condition used the meta-talk strategy

significantly more during the posttest dialogue compared to the participants in the comparison condition, $t(22.06) = 2.26$, $p = 0.03$ (Note: Levene's Test for Equality of Variances was violated, $F_{Posttest\ Dialogue} = 15.74$, $p < .001$). Post-hoc pairwise comparisons revealed that participants in the experimental condition used meta-talk significantly more during the posttest dialogue compared to the initial dialogue, $t(25) = 2.81$, $p = 0.01$. There was a non-significant difference between the mean frequency of use of meta-talk in the comparison condition, $t(37) = 1.36$, $p = 0.18$.

The proportion of participants in the experimental condition who displayed the meta-talk subtype at least once was 4% (n = 1) at initial time and 57.1% (n = 12) at final time, compared to 15.6% (n = 5) at initial time and 46.2% (n = 12) at final time in the comparison condition. These differences were non-significant, $X^2_{Initial\ Dialogue}(1, N = 57) = 2.01$, $p = 0.22$; $X^2_{Posttest\ Dialogue}(1, N = 47) = 0.56$, $p = 0.45$.

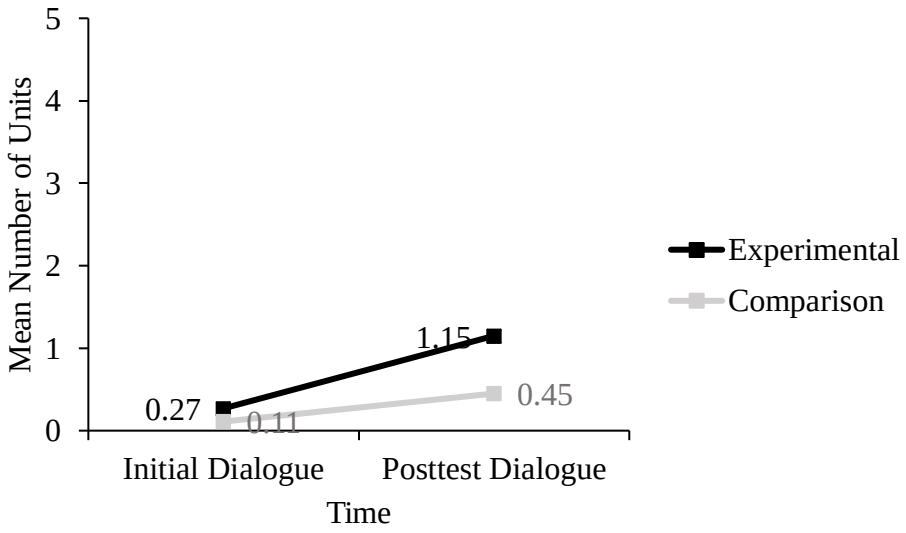
Substantiating. Participants in the experimental condition showed a mean frequency of use of substantiating similar to that of the participants in the comparison condition during the initial dialogue ($M_{Experimental} = 0.27$, $SD_{Experimental} = 0.72$; $M_{Comparison} = 0.11$, $SD_{Comparison} = 0.76$). Usage increased over time in both conditions, with a mean of 1.15 ($SD = 1.85$) in the experimental condition and 0.45 ($SD = 1.52$) in the comparison condition, yielding a significant effect of time, $F(1, 62) = 8.61$, $p = 0.01$, with a non-significant effect of condition, $F(1, 62) = 2.98$, $p = 0.09$ and a non-significant interaction, $F(1, 62) = 1.69$, $p = 0.20$ (see Figure 4).

A planned pairwise comparisons using the Bonferroni corrected alpha level of 0.025 was non-significant, $t(45) = 1.23$, $p = 0.23$. Post-hoc pairwise comparisons revealed that participants in the experimental condition used evidence significantly more during the posttest dialogue compared to the initial dialogue, $t(25) = 2.48$, $p = 0.02$. There was a non-significant difference

between the mean frequency of use of evidence in the comparison condition, $t(37) = 1.40$, $p = 0.17$.

Figure 4

Evidence-Based Statements Usage by Time and Condition



The proportion of participants in the experimental condition who displayed the substantiating subtype at least once was 24% (n = 6) at initial time and 66.7% (n = 14) at final time, compared to 21.9% (n = 7) at initial time and 50% (n = 13) at final time in the comparison condition. These differences were non-significant, $X^2_{Initial\ Dialogue}(1, N = 57) = 0.04$, $p = 0.85$; $X^2_{Posttest\ Dialogue}(1, N = 47) = 1.32$, $p = 0.25$.

Countering. Use of the more powerful counterargument strategies, counter-critique and counter-undermine, showed differing patterns. For counter-critique, there was a significant effect of time, $F(1, 62) = 20.03$, $p < .001$, but no effect of condition, $F(1, 62) = 1.27$, $p = 0.27$, nor an interaction, $F(1, 62) = 0.54$, $p = 0.47$ (see Figure 5a).

A planned pairwise comparisons using the Bonferroni corrected alpha level of 0.025 was non-significant, $t(45) = 0.26$, $p = 0.80$. Post-hoc pairwise comparisons revealed that participants

in the experimental condition used counter-critique significantly more during the posttest dialogue compared to the initial dialogue, $t(25) = 3.52$, $p < .001$. There was also a significant difference between the mean frequency of use of counter-critique in the comparison condition, $t(37) = 2.86$, $p = 0.01$.

The proportion of participants in the experimental condition who displayed the counter-critique subtype at least once was 28% (n = 7) at initial time and 81% (n = 17) at final time, compared to 28.1% (n = 9) at initial time and 88.5% (n = 23) at final time in the comparison condition. These differences were non-significant, $X^2_{Initial\ Dialogue}(1, N = 57) = 0.00$, $p = 1.00$; $X^2_{Posttest\ Dialogue}(1, N = 47) = 0.52$, $p = 0.68$.

For counter-undermine, there was a significant interaction between time and condition, $F(1, 62) = 10.15$, $p < .001$, as well as a significant effect of time, $F(1, 62) = 70.52$, $p < .001$, and condition, $F(1, 62) = 14.15$, $p < .001$ (see Figure 5b).

A planned pairwise comparison using the Bonferroni corrected alpha level of 0.025 revealed that participants in the experimental condition used counter-undermine significantly more during the posttest dialogue compared to the participants in the comparison condition, $t(45) = 3.57$, $p < .001$. Post-hoc pairwise comparisons revealed that participants in the experimental condition used the counter-undermine strategy significantly more during the posttest dialogue compared to the initial dialogue, $t(25) = 4.87$, $p < .001$. There was also a significant difference between the mean frequency of use of counter-undermine in the comparison condition, $t(37) = 2.43$, $p = 0.02$.

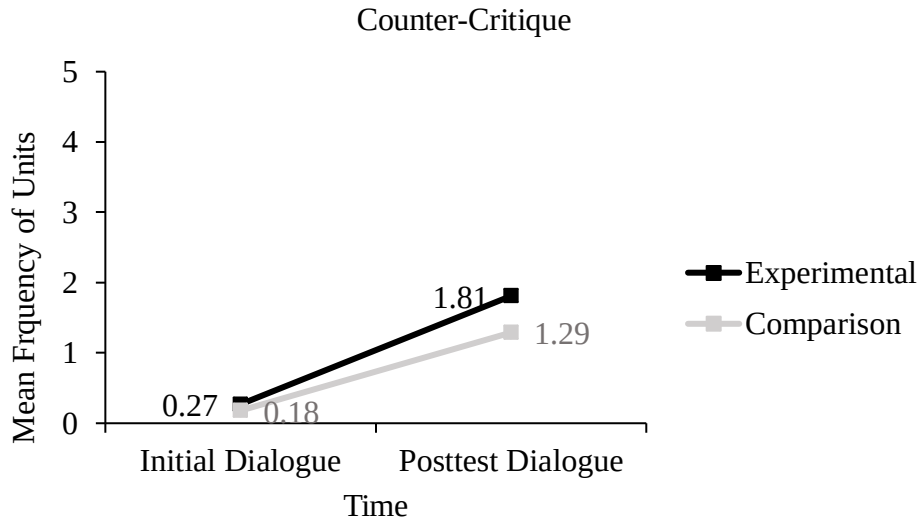
The proportion of participants in the experimental condition who displayed the counter-undermine subtype at least once was 12% (n = 3) at initial time and 100% (n = 21) at final time, compared to 9.4% (n = 3) at initial time and 65.4% (n = 17) at final time in the comparison

condition. This difference reached significance at final time: $X^2_{Initial\ Dialogue}(1, N = 57) = 0.10$, $p = 1.00$; $X^2_{Posttest\ Dialogue}(1, N = 47) = 8.99$, $p < 001$.

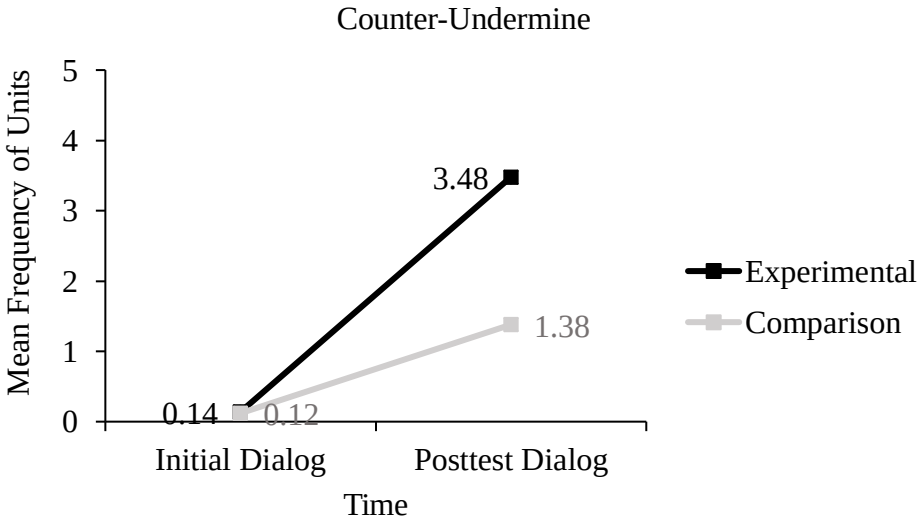
Figure 5

Stronger Counterargument Strategies Usage by Time and Condition

a.



b.



Use of weaker counterargument strategies, counter-alternative and counter-disagree, showed similar patterns. For counter-alternative, the interaction was non-significant, $F(1, 62) = 0.05$, $p = 0.82$, as was the effect of time, $F(1, 62) = 0.02$, $p = 0.89$, and condition, $F(1, 62) = 1.37$, $p = 0.25$, although mean frequency of use in the experimental condition rose to a lesser extent from initial to posttest dialogues than in the comparison condition, as indicated by the mean differences: $M_{Experimental} = 0.15$, $M_{Comparison} = 0.03$.

A planned pairwise comparisons using the Bonferroni corrected alpha level of 0.025 was non-significant, $t(45) = 1.00$, $p = 0.93$. Post-hoc pairwise comparisons revealed that participants in the experimental condition did not use the counter-alternative strategy significantly more during the final dialogue compared to the initial dialogue, $t(25) = 0.26$, $p = 0.80$. There was also a non-significant difference between the mean frequency of use of counter-alternative in the comparison condition, $t(37) = 0.06$, $p = 0.95$.

The proportion of participants in the experimental condition who displayed counter-alternative at least once was 52% (n = 13) at initial time and 65.6% (n = 21) at final time, compared to 66.7% (n = 14) at initial time and 76.9% (n = 20) at final time in the comparison condition. These differences were non-significant, $X^2_{Initial\ Dialogue}(1, N = 57) = 1.08$, $p = 0.30$; $X^2_{Posttest\ Dialogue}(1, N = 47) = 0.61$, $p = 0.44$.

For counter-disagree, there was a non-significant interaction, $F(1, 62) = 0.24$, $p = 0.88$. In both conditions, participants used counter-disagree slightly less during the posttest dialogue ($M_{Experimental} = 0.08$, $SD_{Experimental} = 0.80$; $M_{Comparison} = 0.00$, $SD_{Comparison} = 0.87$) than during the initial dialogue ($M_{Experimental} = 0.19$, $SD_{Experimental} = 0.57$; $M_{Comparison} = 0.16$, $SD_{Comparison} = 0.75$), though the effect of time was non-significant, $F(1, 62) = 1.00$, $p = 0.32$, as was the effect of condition, $F(1, 62) = 0.16$, $p = 0.69$.

A planned pairwise comparisons using the Bonferroni corrected alpha level of 0.025 was non-significant, $t(45) = 0.67$, $p = 0.51$. Post-hoc pairwise comparisons revealed that participants in the experimental condition did not use the counter-disagree strategy significantly more during the final dialogue compared to the initial dialogue, $t(25) = 0.77$, $p = 0.45$. There was also a non-significant difference between the mean frequency of use of counter-disagree in the comparison condition, $t(37) = 0.78$, $p = 0.44$.

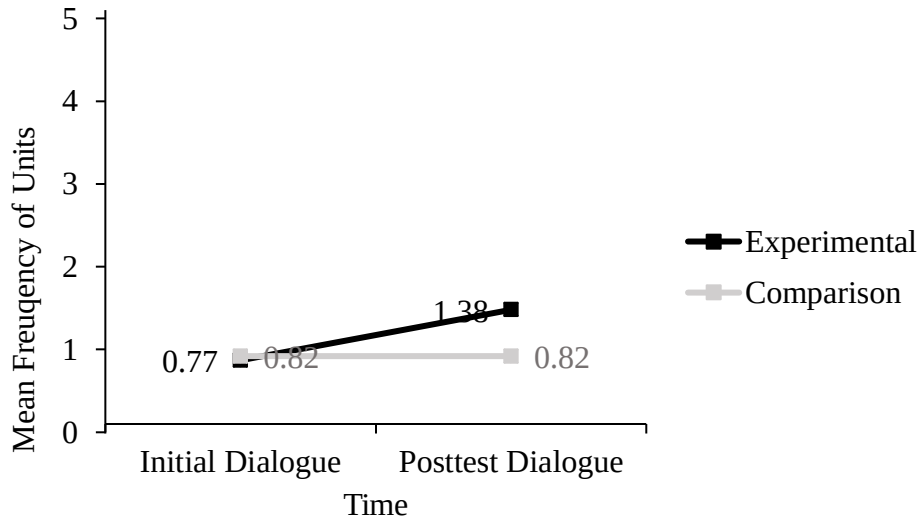
The proportion of participants in the experimental condition who displayed the counter-disagree type at least once was 20% (n = 5) at initial time and 23.8% (n = 5) at final time, compared to 31.3% (n = 10) at initial time and 38.5% (n = 10) at final time in the comparison condition. These differences were non-significant, $X^2_{Initial\ Dialogue}(1, N = 57) = 0.92$, $p = 0.34$; $X^2_{Posttest\ Dialogue}(1, N = 47) = 1.15$, $p = 0.28$.

Concession. The two subtypes, "support-other" and "weaken-own", were analyzed in combination. Mean frequency of use increased from initial to posttest times in the experimental condition, but did not yield a significant effect of time, $F(1, 62) = 0.19$, $p = 0.67$ (see Figure 6). The effect of condition was not significant, $F(1, 62) = 0.26$, $p = 0.60$. The interaction was not significant, $F(1, 62) = 2.12$, $p = 0.15$.

A planned pairwise comparisons using the Bonferroni corrected alpha level of 0.025 was non-significant, $t(45) = 0.46$, $p = 0.65$. Post-hoc pairwise comparisons revealed that participants in the experimental condition did not use concession significantly more during the posttest dialogue compared to the initial dialogue, $t(25) = 1.14$, $p = 0.26$. There was also no significant difference between the mean frequency of use of concession in the comparison condition, $t(37) = 0.00$, $p = 1.00$.

Figure 6

Concession Statements Usage by Time and Condition



The proportion of participants in the experimental condition who displayed the concession type at least once was 60% (n = 15) at initial time and 66.7% (n = 14) at final time, compared to 65.6% (n = 21) at initial time and 69.2% (n = 18) at final time in the comparison condition. These differences were non-significant, $X^2_{Initial\ Dialogue}(1, N = 57) = 0.19$, $p = 0.66$; $X^2_{Posttest\ Dialogue}(1, N = 47) = 2.51$, $p = 0.11$.

Essay Analysis

The essays chosen for comparison across condition were participants individual essays for the animal research topic, the last topic on which participants wrote topic essays within the workshop, and the posttest essays on the transfer topic (not debated within the workshop).

The unit of analysis consisted of an idea unit within an essay. These were further categorized into four functional types based on how they served the writer's claim, *support-own*, *weaken-own*, *support-other*, *weaken-other* (see Table 2).

Table 2

Examples of Functional Idea Units in Analytic Scheme for Coding Individual Essays

Argumentive Function	Example
Support my-own	Testing upon animals achieves and gets prudent results (*pro animal research*) Alternatives allow researchers to make predictions (*pro alternative methods*)
Weaken my-own	A very major fault in testing on animals is that animals are put in danger (*pro animal research*) These alternative methods are still in development (*pro alternative methods*)
Support-Other	There are other ways to find cures and vaccines (*pro animal research*) Animal testing may be the only option sometimes (*pro alternative methods*)
Weaken-Other	Computers might not be as accurate as testing on animals (*pro animal research*) There needs to be a lot of precautions in place to make sure the animal is well cared for (*pro alternative methods*)

In addition, two adjacent statements, one occurring immediately after the other and explicitly connected to one another, typically by conjunctions “however”, “but”, or “although”, were categorized into *However* arguments. Possible combinations include support-other with support-own; support-other with weaken-other; weaken-own with weaken-other; and weaken-own with support-own.

The statistical analysis consisted of a negative binomial regression. The condition was the between-subjects two-level factor. We chose this method because our dependent variable(s)

consist of only non-negative integer values and the variance of the dependent variable is greater than the mean. The dependent variable is substantially positively skewed. Given the non-normal and highly skewed nature of the dependent variable, we log-transformed the outcome variables, but this led to difficulties in the interpretation of the estimates which may have produced misleading conclusions. Standard regression techniques (such as ordinary least squares regression) are not suitable for this data set (Hilbe, 2011a; Hilbe, 2011b; NCSS Statistical Software, 2021; Ver Hoef & Boveng, 2007). Therefore, Poisson regression is the first-choice modeling technique, but the mean and the variance of the Poisson distribution are equal to the parameter estimate of the expected frequency for the response variable (Hilbe, 2011c; Ver Hoef & Boveng, 2007) (see appendix). However, the assumption of identical mean and variance was not satisfied for the data set (see appendix). The greater ratio of variance to mean leads to overdispersion frequently caused by heterogeneity among observations (Hilbe, 2011b). We therefore used a negative binomial regression to overcome the problem of overdispersion.

Negative binomial regression includes a random error term to relax the Poisson regression assumption of identical mean and variance, giving the explanatory variable(s) more predictive power. In analyzing the subset of dependent variables, we considered the model below to model the differential frequency of use of the respective dependent variable for participants in the experimental and comparison conditions:

$$\exp(\ln Y_i) = \beta_0 + \beta_1 X_{1i}$$

where β_0 is the intercept ($X \equiv 1$); Y_i is the average frequency of the response variable of interest in the individual animal research essays for participant *i*; X_{1i} is the predictor Condition, a dichotomous dummy variable coded "1" for experimental condition and "0" as the comparison

condition (so that the comparison condition serves as the reference group) for participant *i;* $\beta_1$is the parameter to be estimated.

An additional analysis of proportion of use, i.e., the proportion of participants who used a particular type once or more, was conducted. Pearson's Chi-square Test(s) of Independence were conducted to explore these relationships. A Fisher's Exact Test of Independence was used in cases where the Pearson's Chi-square Test of Independence could not be used due to low expected counts in two-by-two tables. For the purposes of failing to reject or rejecting the two-tailed null hypothesis, the significance level was set at an alpha level of 0.05.

Of the 57 of participants who completed the individual animal research essays, 26 (46%) were in the experimental condition and 31 (54%) were in the comparison condition. Of the 53 participants who completed the individual posttest essays on the transfer topic, 23 (43%) were in the experimental condition and 30 (57%) were in the comparison condition.

Identifying and Coding of Essays

All participant identifying information was removed prior to coding to preserve privacy and minimize bias. Essays were coded by the author and a research assistant who was familiar with the study but blind to experimental condition. Essays were first segmented into idea units. Acceptable inter-rater agreement of 91% for segmenting was achieved among two raters on 25% of the data, Cohen's kappa = 0.86, $p < .001$.

For coding of idea units into functional types, percent agreement between the same two coders was 86%, Cohen's kappa (Cohen, 1960) was acceptable ($k = 0.83$, $p < .001$). Disagreements were resolved through discussion. The remaining essays were coded by the author.

Idea units that consisted of an assertion with no reason or evidence to support it were not included in the analysis of types. Nor were those whose meanings were not clear or discernible, or were repetitions of a previous idea unit without elaboration.

Performance on Individual Essay on Animal Research by Condition

Number of Idea Units. The estimates of the model parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) are reported in Table 3.

Table 3

Estimation Results of the Negative Binomial Regression on Animal Research Essay Idea Units

$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
Intercept	1.94	0.14	(1.66 to 2.21)	6.94
Condition[a]	0.16	0.20	(-0.24 to 0.55)	1.17

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

The estimated coefficients of the predictor variable were not significant. In this model, the mean number of idea units in the experimental condition was 1.17 times (i.e., 17% higher) compared to that of the participants in the comparison condition (IRR = 1.17, p = 0.44).

Functional Types. A negative binomial regression was conducted on each of the four functional subtypes to determine if there were differences in group means. The estimates of the model parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) are reported in Table 4.

The estimated coefficients of the predictor variable were not statistically significant for each of the four functional types. Participants in the experimental condition were found to be

more likely to have higher numbers of support-own (IRR = 1.13, $p = 0.54$), weaken-own (IRR = 1.84, $p = 0.20$), and support-other (IRR = 2.39, $p = 0.10$) counts than participants in the comparison condition. Although three of the four types were in the expected direction of superiority of the experimental group, these differences did not reach statistical significance.

Table 4

Estimation Results of the Negative Binomial Regression on Animal Research Essay Functional Types

Response Variable	$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
Support-Own	Intercept	1.45	0.14	(1.18 to 1.72)	4.26
	Condition[a]	0.12	0.20	(-0.27 to 0.51)	1.13
Weaken-Own	Intercept	-1.04	0.36	(-1.74 to -0.34)	0.36
	Condition[a]	0.61	0.48	(-0.33 to 1.55)	1.84
Support-Other	Intercept	-1.24	0.40	(-2.03 to -0.45)	0.29
	Condition[a]	0.87	0.53	(-0.17 to 1.91)	2.39
Weaken-Other	Intercept	0.71	0.21	(0.30 to 1.12)	2.03
	Condition[a]	-0.04	0.31	(-0.65 to 0.58)	0.97

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

Analysis of how many participants ever made use of the other three types in their essays showed weaken-other to be the most frequently used, after support-own, which appeared in virtually all essays. In particular, participants in the experimental condition used weaken-other 65.4% (n = 17) and support-own 100% (n = 26) compared to 71% (n = 22) and 93.5% (n = 29), respectively, in the comparison condition. Condition differences with respect to whether participants ever showed each type were not significant: $X^2_{Support\text{-}own}(1, N = 57) = 1.74$, $p = 0.50$; $X^2_{Weaken\text{-}own}(1, N = 57) = 1.70$, $p = 0.25$; $X^2_{Support\text{-}other}(1, N = 57) = 1.01$, $p = 0.38$; $X^2_{Weaken\text{-}other}(1, N = 57) = 0.20$, $p = 0.78$.

Nor did these differences reach significance when the two belief-congruent types (support-own and weaken-other) were combined, and the two belief-incongruent types (weaken-own, support-other) were combined. Table 5 shows the estimates of the model parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) for belief-statements.

Table 5

Estimation Results of the Negative Binomial Regression on Animal Research Essay Belief-Statements

Response Variable	$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
Belief-Congruent	Intercept	0.50	0.14	(0.22 to 0.77)	1.65
	Condition[a]	0.01	0.21	(-0.40 to 0.41)	1.01
Belief-Incongruent	Intercept	-0.80	0.27	(-1.32 to -0.27)	0.45
	Condition[a]	0.48	0.35	(-0.21 to 1.17)	1.62

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

The estimated coefficients of the predictor variable were not significant for each of the belief-statement types. Participants in the experimental condition were found to be more likely to have roughly similar numbers of belief-congruent statements (IRR = 1.01, p = 0.99) compared to participants in the comparison condition and had higher numbers of belief-incongruent statements (IRR = 1.62, p = 0.17) counts than participants in the comparison condition.

Did conditions differ with respect to whether participants ever showed each type? Percentages of participants ever showing belief-congruent statements were 100% (n = 26) in the experimental condition and 96.8% (n = 30) in the comparison condition. Fisher's Exact Test on belief-congruent statements yielded a non-significant result, $X^2_{Belief\text{-}congruent}(1, N = 57) = 0.85$, p =

1.00. Percentages of participants ever showing belief-incongruent statements were 61.5% (n = 16) in the experimental condition and 35.5% (n = 11) in the comparison condition. A Pearson's Chi-Square Test of Independence was conducted on belief-incongruent statements. The result was non-significant, $X^2_{Belief\text{-}incongruent}(1, N = 57) = 3.85$, $p = 0.07$.

Evidence-Based Units. A negative binomial regression was conducted on the mean number of all evidence-based statements (i.e., personal + shared evidence), along with a negative binomial regression on the mean number of shared evidence-based statements. (Note: Evidence based units were re-counted based on whether the functional use of the unit did/did not include evidence). The estimates of the model parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) are reported in Table 6.

Table 6

Estimation Results of the Negative Binomial Regression on Animal Research Essay Evidence-Based Statements

Response Variable	$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
Total Evidence	Intercept	1.01	0.22	(0.57 to 1.45)	2.74
	Condition[a]	0.10	0.33	(-0.55 to 0.75)	1.11
Shared Evidence	Intercept	-0.44	0.30	(-1.03 to 0.15)	0.65
	Condition[a]	0.58	0.42	(-0.23 to 1.40)	1.79

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

The estimated coefficients of the predictor variable were not statistically significant for each of the evidence-based statements. Participants in the experimental condition were found to be more likely to have higher numbers of total evidence-based statement (IRR = 1.11, $p = 0.76$)

and shared evidence-based statement (IRR = 1.79, $p = 0.16$) counts than participants in the comparison condition.

Pearson's Chi-square Tests of Independence were conducted for proportion of participants ever including evidence-based statements. Percentages of participants ever showing evidence-based statements were 80.8% (n = 21) in the experimental condition and 58.1% (n = 18) in the comparison condition. Percentages of participants ever showing shared evidence-based statements were 53.8% (n = 14) in the experimental condition and 32.3% (n = 10) in the comparison condition. Condition differences were non-significant for total and for shared evidence: $X^2_{Total\ Evidence}(1, N = 57) = 3.37$, $p = 0.09$; $X^2_{Shared\ Evidence}(1, N = 57) = 2.70$, $p = 0.12$.

However Arguments. A negative binomial regression was conducted on the total number of however arguments to determine if there were differences in group means. However arguments consist of adjacent non-congruent codes that have the "this, but that" structure, e.g., support-other and weaken-other; support own and weaken own; support-other and support-own; weaken-other and weaken-own (note that each however argument can function equally as well the other way around, as long as they have the however argument structure). Because there were few counts of the distinct however arguments, we only analyze the total count here. The mean number of total however arguments in the experimental condition was 1.15 ($SD = 1.29$) in the experimental condition and 1.06 ($SD = 1.69$) in the comparison condition. The estimates of the model parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) are reported in Table 7.

The estimated coefficients of the predictor variable was not significant for however arguments. Participants in the experimental condition were found to be more likely to have

similar numbers of total however arguments (IRR = 1.01, $p = 0.96$) counts than participants in the comparison condition.

Table 7

Estimation Results of the Negative Binomial Regression on Animal Research However Arguments

Response Variable	$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
However Arguments	Intercept	0.13	0.20	(-0.26 to 0.52)	1.14
	Condition[a]	0.96	0.29	(-0.55 to 0.58)	1.01

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

Pearson's Chi-square Tests of Independence were conducted for proportion of participants ever including however arguments. Percentages of participants ever showing however arguments were 50% (n = 13) in the experimental condition and 41.9% (n = 13) in the comparison condition. Condition differences were non-significant for however arguments: X^2 (1, N = 57) = 0.15, p = 0.79.

Performance on Individual Essay on the Transfer Topic by Condition

Number of Idea Units. The estimates of the model parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) are reported in Table 8.

The estimated coefficients of the predictor variable were not statistically significant. In this model, the mean number of idea units in the experimental condition was 1.47 times (i.e., 47% higher) compared to that of the participants in the comparison condition (IRR = 1.47, p = 0.08).

Table 8

Estimation Results of the Negative Binomial Regression on Transfer Topic Essay Idea Units

$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
Intercept	1.09	0.15	(0.80 to 1.38)	2.97
Condition[a]	0.38	0.22	(-0.04 to 0.81)	1.47

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

Functional Types. A negative binomial regression was conducted on each of the four functional subtypes to determine if there were differences in group means. The estimates of the model parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) are reported in Table 9.

Table 9

Estimation Results of the Negative Binomial Regression on Transfer Topic Essay Functional Types

Response Variable	$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
Support-Own	Intercept	0.73	0.15	(0.44 to 1.01)	2.07
	Condition[a]	0.03	0.22	(-0.40 to 0.47)	1.03
Weaken-Own	Intercept	-3.40	1.00	(-5.36 to -1.44)	0.03
	Condition[a]	1.36	1.15	(-0.90 to 3.63)	3.91
Support-Other	Intercept	-1.61	0.45	(-2.49 to -0.73)	0.20
	Condition[a]	0.67	0.60	(-0.51 to 1.85)	1.96
Weaken-Other	Intercept	-0.41	0.29	(-0.98 to 0.16)	0.67
	Condition[a]	0.93*	0.40	(0.17 to 1.71)	2.54

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

The estimated coefficients of the predictor variable were not statistically significant for three of the four functional types (support-own, weaken-own, support-other), but was significant for weaken-other. Participants in the experimental condition were found to be more likely to have roughly similar numbers of support-own (IRR = 1.03, $p = 0.89$) statements than participants in the comparison condition; higher numbers of weaken-own (IRR = 3.91, $p = 0.24$), support-other (IRR = 1.96, $p = 0.27$), and weaken-other (IRR = 2.54, $p = 0.02$) counts than participants in the comparison condition. Although all of the four types were in the expected direction of superiority of the experimental group, only one of these differences reached significance.

Analysis of how many participants ever made use of the other three types in their essays showed weaken-other to be the most frequently used, after support-own, which appeared in the large majority of essays (91% and 80% in experimental and comparison conditions respectively). In the experimental condition weaken-other was used by 60.9% (n = 14) of participants in the experimental condition and 43.3% (n = 13) in the comparison condition. Support-own was used by 100% (n = 26) of participants in the experimental condition and 93.5% (n = 29) in the comparison condition. In the experimental condition support-other was used by 34.6% (n = 9) of participants in the experimental condition and 22.6% (n = 7) in the comparison condition, and weaken-own was used by 13% (n = 3) of participants in the experimental condition and 3.3% (n = 1) in the comparison condition. Condition differences with respect to how many participants ever made use of the four functional types were not significant: $X^2_{Support\text{-}own}(1, N = 53) = 1.30$, $p = 0.44$; $X^2_{Weaken\text{-}own}(1, N = 53) = 1.76$, $p = 0.31$; $X^2_{Support\text{-}other}(1, N = 53) = 2.32$, $p = 0.18$; $X^2_{Weaken\text{-}other}(1, N = 53) = 1.60$, $p = 0.27$.

Nor did these differences reach significance when the two belief-congruent types (support-own and weaken-other) were combined. Table 10 shows the estimates of the model

parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) for belief-statements.

Table 10

Estimation Results of the Negative Binomial Regression on Transfer Topic Essay Belief-Statements

Response Variable	$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
Belief-Congruent	Intercept	1.01	0.15	(0.71 to 1.30)	2.73
	Condition[a]	0.34	0.22	(-0.09 to 0.77)	1.40
Belief-Incongruent	Intercept	-1.46	0.42	(-2.29 to -0.62)	0.23
	Condition[a]	0.81	0.56	(-0.29 to 1.90)	2.24

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

The estimated coefficients of the predictor variable were not significant for each of the belief-statement types. Participants in the experimental condition were found to be more likely to higher numbers of belief-congruent (IRR = 1.40, p = 0.12) and belief-incongruent (IRR = 2.24, p = 0.15) counts than participants in the comparison condition. Although both belief-statement types were in the expected direction of superiority of the experimental group, none of these differences reached statistical significance.

Belief-congruent statements appeared in the essays of 61.5% (n = 16) of participants in the experimental condition and 35.5% (n = 11) in the comparison condition. Belief-incongruent statements appeared in the essays of 34.8% (n = 8) of participants in the experimental condition and 16.7% (n = 5) of participants in the comparison condition. Fisher's Exact Tests yielded non-significant differences: $X^2_{Belief\text{-}Congruent}(1, N = 53) = 3.32$, $p = 0.12$; $X^2_{Belief\text{-}Incongruent}(1, N = 53) = 2.31$, $p = 0.20$.

Evidence-Based Units. A negative binomial regression was conducted on the mean number of total evidence-based statements. Because information that could be used as evidence was not shared during the transfer topic essay, evidence-based statements consisted only of personal evidence. The estimates of the model parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) are reported in Table 11.

Table 11

Estimation Results of the Negative Binomial Regression on Transfer Topic Essay Evidence-Based Statements

Response Variable	$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
Total Evidence	Intercept	-0.46	0.56	(-1.55 to 0.64)	0.63
	Condition[a]	0.62	0.83	(-1.01 to 2.24)	1.85
Shared Evidence	Intercept	—	—	—	—
	Condition[a]	—	—	—	—

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

The estimated coefficients of the predictor variable were not statistically significant for total evidence-based statements. Participants in the experimental condition were found to be more likely to have higher numbers of total evidence-based statements (IRR = 1.85, p = 0.46) counts than participants in the comparison condition.

Pearson's Chi-square Tests of Independence were conducted for proportion of use of total evidence-based statements by condition. Total evidence-based statements appeared in 30.4% (n = 7) of the transfer topic essays of the participants in the experimental condition and 16.7% (n =

5) of the essays of the participants in the comparison condition. The difference was non-significant: $X^2_{Total\ Evidence}(1, N = 53) = 1.41, p = 0.32$.

However Arguments. A negative binomial regression was conducted on the total number of however arguments to determine if there were differences in group means. Because there were few counts of the distinct however arguments, we only analyze the total count here. The mean number of total however arguments in the experimental condition was 0.52 (*SD* = 1.08) in the experimental condition and 0.17 (*SD* = 0.46) in the comparison condition. The estimates of the model parameters ($\hat{\beta}$), standard errors (Std Err of $\hat{\beta}$), 95% confidence interval (CI) for the $\hat{\beta}$, and the incidence rate ratios (IRRs) are reported in Table 12.

Table 12

Estimation Results of the Negative Binomial Regression on Transfer Topic However Arguments

Response Variable	$\hat{\beta}$	Estimate	Std Err of $\hat{\beta}$	Conf. Int. of $\hat{\beta}$	IRR
However Arguments	Intercept	-1.72	0.55	(-2.80 to -0.65)	0.18
	Condition[a]	1.07	0.71	(-0.32 to 2.46)	2.92

[a] Condition = 0 is taken as the reference category; IRRs, incident rate ratios

*p <0.05

The estimated coefficients of the predictor variable was not statistically significant for however arguments. Participants in the experimental condition were found to be more likely to have higher numbers of total however arguments (IRR = 2.92, p = 1.07) counts than participants in the comparison condition.

Pearson's Chi-square Tests of Independence were conducted for proportion of participants ever including however arguments. Percentages of participants ever showing however arguments were 26.1% (n = 6) in the experimental condition and 13.3% (n = 4) in the

comparison condition. Condition differences were non-significant for however arguments: X^2 (1, N = 53) = 0.15, p = 0.48.

Chapter 5: Discussion

The purpose of this research was to examine whether young adolescents who interact with an expert arguer demonstrate advancements in argumentive skill and strategy use, both in dialogues and in individual essays. Young adolescents participated in an online adaptation of the AWM dialogic argumentation-based curriculum (described in Kuhn, 2018b and Kuhn et al., 2016a) that promotes deep engagement with peer discourse on a series of topics. The adaptation concentrated on dialogic interaction (i.e., the game phase) in which participants engaged in electronic discourse, anonymously, with a series of rotating peers in private, one-to-one discussion rooms.

Groups differed with respect to the dialogic interactions that took place. Participants in the experimental condition engaged in dialogue sessions with alternating peer opponents and unknowingly, with an adult expert and students in the comparison condition only engaged in dialogue with a peer opponent.

Analysis of the intervention centered around argument strategies, with the two stronger of the four counterargument strategies (counter-critique and counter-undermine) and more sophisticated strategies (meta-talk, evidence, and concession) implying developmental advancements in dialogues. Analysis of essays likewise centered around functional types, identifying advanced argumentive functions as the more sophisticated, weaken-other, support-other, and weaken-own. An additional analysis on participants' ratings of their dialogic partners' on a scale from zero to 100 served as a manipulation check, with higher scores indicating that partners' were good arguers'.

Summary of Results

Overall, the results suggest that dialogic engagement in written form with more capable others, as well as with peers of similar ability, enhances argumentation skill. This finding is consistent with and extends upon recent work by Mayweg-Paus et al. (2015) and Papathomas and Kuhn (2017). The results also suggest that the dialogic process in written form is a path to argumentive competence in expository argumentive writing. This finding is consistent with the developmental progression observed in earlier work that suggests that statements that weaken claims appear to be a more challenging achievement than statements that serve in support of a claim, and statements that support opposing claims or weaken one's own claims require some reconciliation.

Manipulation Check

The results suggest that the participants in the experimental condition perceived the manipulation, i.e., they recognized the superior quality of an expert's, compared to a peer's, contributions to dialogue.

Dialogue Effects

The results support the hypothesis that dialogic engagement with a more capable partner, as well as with peers of similar ability, advances young adolescents argumentation skills. Gains in procedural skills were assessed at initial and final dialogue. Analysis of dialogues suggests that dialogic engagement with a more capable partner (when interaction is anonymous and social and relational status is unknown) contributes to the development of argumentation skill. In particular, both groups improved in use of stronger counterargument strategies (counter-critique, counter-undermine) and sophisticated strategies (questions) overtime, but the experimental group showed greater skill in using the most advanced counterargument strategy (counter-undermine),

along with more sophisticated strategies that reflect meta-strategic understanding (meta-talk, evidence), compared to the comparison group.

Essay Effects

The results support the hypothesis that dialogue is a promising pedagogical path to the development of individual expository argument skill in written form. Gains in individual argument skills were assessed in the last debate-topic essay (animal research [week 3]) and in the non-debate-topic essay (transfer topic [week 4]). Although not all of the essay results referred to herein were statistically significant, analysis of essays suggests that development progresses in the direction of integrative complexity that typically manifests through dual argumentive strategy (support-own and weaken-other) and later through integrative strategies (typically support-other and rarely weaken-own). In particular, participants in both conditions used weaken-other statements most frequently in their animal research and transfer topic essays, besides a substantially greater effect in the transfer topic essay for participants in the experimental group. Sustained engagement in dialogic arguing activities gives novice arguers opportunities to confront their opponents' arguments, and as a result, the biases in thinking focused on what "I say" improve, transitioning to thinking focused on what "my opponents say".

Over time and with extended opportunity and practice, both groups transitioned from the frequent use of dual- to integrative- argumentive strategies, but the experimental group showed greater skill gains in individual expository writing. In particular, the participants in the experimental group used more belief-incongruent statements, along with supporting statements in their animal research essays, but more belief-incongruent statements in their transfer topic essays than the comparison group. Likewise, the experimental group used more statements that support the opposing position in their animal research essays and in the transfer topic essays, in

addition to using more statements that weakened their own position compared to the comparison condition. We also found that the "however" structure made its way into students essays.

The results suggest engagement dialogically in writing left both groups with an enhanced meta-level understanding of the role of evidence in argument, but the experimental group showed greater skill in coordinating evidence with claims. In particular, in both essays, participants in the experimental group used more evidence than the participants in the comparison condition. The experimental group also used more shared evidence in the animal research essays relative to the comparison condition, a finding that is consistent with prior research (Kuhn & Moore, 2015; Macagno, 2016). Evidence in argument is an indicator of argument skill and an important and relevant measure to address the research question in the present study. To address shared evidence the partner needs to coordinate the claim with the evidence (Hemberger et al., 2017; Kuhn et al., 2016). Using evidence in argument is powerful because it strengthens or weakens a claim and, once shared, cannot be ignored (Kuhn et al., 2016b). The experts employed evidence or requests for empirical justification on a regular basis, which the participants in the experimental condition apparently adopted and integrated into their dialogues.

Limitations and Future Research

With respect to the study design, it was not logistically possible to control for partial contamination of treatment effects as participants in the experimental condition engaged in argumentive discourse with participants in the comparison condition (note that participants in the comparison condition rarely, if ever, interacted with an expert dialogic partner). This was, in part, due to the small sample size, though previous research in classroom contexts has effectively controlled for contamination of treatment effects with relatively smaller sample sizes (e.g.,

Mayweg-Paus et al., 2015; Papathomas & Kuhn, 2017). It may be easier to control for contamination of treatment effects in formal classroom contexts where students are typically assigned a classroom, although formal classroom contexts may bring about additional unforeseen challenges. Still, given that we found significant effects when the effects of contamination typically dull or mask existing relationship(s) between independent and dependent variable(s), is impressive, and suggests that contamination may have been minimized. Regardless, future researchers interested in investigating interaction characteristics using the online adaptation of AWM ought to carefully consider logistical design challenges that threaten internal validity, and should investigate larger, more representative samples.

An additional procedural drawback of the design was that we were not able to investigate whether participants used shared versus personal information Q&As during the initial and final dialogue assessments. Although lack of this data does not preclude answering our main research question, it does, in some ways, artificially restrict researchers' range of understandings of evidence use in dialogues and in essays. Future empirical investigations would benefit from altering the procedural aspects, perhaps by adding additional assessments or adjusting assessment administration timelines.

A limitation that threatens the study's validity is that the study was conducted during the COVID-19 pandemic of 2020. An added threat to external validity is the fact that we were unable to collect comprehensive data on sample demographics. Taken together, it may not be possible to generalize findings. However, given that the pandemic-related changes to teaching and learning continue to last as of our writing of this paper (1+ years post pandemic), it is possible that the findings may be valid only in certain contexts or in times of crisis (although tentative).

Theoretical Implications

On the theoretical front, the study contributes support for mechanisms grounded in constructivist theories of development. Vygotsky describes a developmental mechanism by which students learn from the social to the individual through processes of internalization (or interiorization; Kuhn, 2018a; Kuhn, 2019; Kuhn et al., 2016b) in a zone of proximal development. The effect of the experimental manipulation follows the theoretical pattern of proximal development, with advanced argumentation skills developing for participants engaging with an expert who models strategies, beyond the participants' demonstrated individual capacity in the social context of electronic argumentive discourse (meaning that modeling occurs exclusively through language where in-person cues are absent). In the present study, arguers are faced with the need to contribute relevantly to the dialogue when interacting electronically with an interlocuter who displays deeper and more effective argumentive moves (an expert) and need to adapt their communicative behavior to adopt the sophisticated strategies and moves made by the more skilled opponent. This effect made its way into the individual essay context and did so to a greater extent for the participants in the experimental condition than comparison condition, supporting earlier claims that apprenticeship learning is a potentially powerful mechanism for developing argumentive discourse skill.

Our findings are consistent with earlier studies (Mayweg-Paus et al., 2015; Papathomas & Kuhn, 2017) that showed that the benefits of peer collaboration extend to equal- and unequal-ability peers, which suggests that the social is not a unidirectional source of influence on individual learning and development. In fact, Kolikant and Pollack (2015) remind us that development also occurs from the individual to social. Although social and cognitive development are not explicitly distinct entities, the extant literature on dialogic education and

argumentation have focused on cognitive aspects of development where epistemological issues are considered more heavily than relational issues. Because our findings, on the one hand, support development from the social to individual, but on the other, support the idea that there is more than one direction of influence, we think our findings may be explained equally well by more than one theory and in particular, consider theory from the social domain; specifically, sociocognitive conflict and conflict regulation.

Theoretical Considerations of Development from the Social Domain

Sociocogntive conflict and conflict regulation theory possibly lends insight into the mediating processes and effects of relational variables in competence relevant situations; specifically, quantitative and qualitative aspects of social comparison (i.e., social status among and relationships with interlocutors) and their influence on the outcome of the interaction (i.e., the influence relationship) (Butera & Mugny, 1995; Darnon et al., 2007; Quiamzade et al., 2003). Social comparison functions as a means of obtaining information that is useful in assessing relative competence, along with relative knowledge and understanding (Butera & Darnon, 2017; Butera et al., 2019; Jordan-O'Reilly, 2013; Križan & Gibbons, 2014). Although speculative, the anonymous dialogic context in the current study may have made it difficult for participants to collect relational information, e.g., partner status (e.g., teacher, peer, etc.) ability (e.g., expert, novice, etc.; although participants assigned higher ratings to expert partners relative to peer partners), and connectedness (e.g., stranger, friend, etc.), possibly suppressing social comparison mechanisms and influencing learning and development. Nevertheless, social comparison has dual (interpersonal) and cognitive (intrapersonal) functions and is a mediating process involved in conflict regulation, which concerns the relations between the extent to which interpersonal

(social, interactional) conflict and the extent to intrapersonal (individual, cognitive) competencies develop (Butera & Darnon, 2017; Butera et al., 2019).

Social influence is important to consider because when individuals are uncertain of knowledge and competences, e.g., their own achievement-related attributes (i.e., beliefs, attitudes, values, knowledge, etc.), they are more likely to be influenced by others in the immediate social context (Lun et al., 2007). This draws attention to the quantitative and qualitative aspects of social comparison influence dynamics that predict manifest (i.e., immediate) and latent learning when low-competence targets interact with high-competence sources (as in the experimental condition in the current study and previously, Mayweg-Paus et al. (2015) and Papathomas and Kuhn (2017)), and latent learning when low-competence targets and sources interact (as in the comparison condition in the current study and previously, Mayweg-Paus et al. (2015) and Papathomas and Kuhn (2017)) (Butera & Darnon, 2017). In addition to the level of learning, the type of learning and development achieved, i.e., the influence relationship, is internalization in the former condition, and experimentation and exploration in the latter condition; both induce deep learning (as opposed to superficial learning) (see Table 13).

Table 13

Sociocognitive Conflict and Conflict Regulation Processes and Outcomes

Influence dynamics		**Level of learning**		**Social influence**		**Characteristics**
Target	Source	Manifest	Latent	Superficial	Deep	
Low	High	X	X		X	• Internalization • Compliance
Low	Low		X		X	• Experimentation • Exploration, discovery

Again, although speculative, influence relationship may explain why the current and earlier studies found immediate effects of learning in the experimental condition. Future research can

extend upon those prior by adding measures to assess both immediate and latent learning, accounting for relational variables.

That said, however, while it appears that current and earlier study findings may be hypothetically consistent with sociocognitive conflict and conflict regulation theory, Asterhan (2018) reminds us that sociocognitive conflict and regulation studies place students in well-structured situations where there is a clearly-defined answer (or answers), unlike argumentation situations where there is no clear-cut answer(s) to ill-structured complex, socio-scientific issues. Additionally, in an argumentation context, viewpoints and ideas may not conflict per se and therefore, the interlocuters must weigh the plausibility of the alternative perspectives exchanged in the discourse. Interlocuters may not initially perceive compliance as necessarily beneficial when disagreement(s) encountered in dialogue raises doubts in the validity of both source's and target's knowledge state or status, and relative competence (Butera et al., 2019). Simply put, researchers ought not generalize findings from studies involving sociocognitive conflict contexts and regulation processes to argumentation contexts. That is not to say, however, that it is not worth thinking analogously about these phenomena and concepts. Sociocognitive conflict and argumentive discourse are, in part, defined and are defined by, meanings attributed to the context-bound and situation-specific interactions and therefore, scholars advocate empirical investigations of regulation processes—as well as argumentive reasoning and skill development—in the contexts within which understanding is acquired. In fact, this is the central justifying assumption for empirical investigations of dialogic education, and in fact, within the past decade, research has begun to specifically address relational variables characterizing argumentation contexts.

Recent work by Asterhan and Babichenko (2015) on interaction characteristics showed that undergraduate students who believed they were interacting with a computer agent gained in argumentation skills compared to students who believed they were interacting with a peer of equal status and ability (when really, both groups were interacting with a research confederate). Although speculative, interacting with a computer agent limits potential negative social threats and students, therefore, may not only feel freer to share ideas they otherwise may not have, but also be more open to criticism (Asterhan & Babichenko, 2015). Anonymity possibly yields similar dialogic experiences (Ainsworth et al., 2011). Moreover, students may be more likely to adopt information (whether correct or incorrect) given that computer agents are often perceived as authoritative (Asterhan & Babichenko, 2015). Similarly, and again, although speculative, students may be more likely to adopt more sophisticated argumentation moves of more capable dialogic partners when relational status is unknown, but the other partner competency is perceived as more capable.

To summarize this section, we propose a closer examination of interpersonal, affective, and motivational features of dialogic activities characterizing the argumentation context that may help explain why and how dialogic approaches to pedagogy yield gradual improvements in argumentive reasoning and development. In particular, researchers may want to specifically explore intricately related phenomenon such as epistemic or relational regulation, and threat appraisal, along with the structure of achievement-related contexts. Future research may not only shed light on developmental mechanisms, mediators, etc., but also may pave the way for altering how we think about complex psychosocial phenomena.

Novel Contributions and Distinctions for Research

Findings from the current study expand upon our understandings of dialogic contexts in several key ways. The fact that participants argumentive competencies improved when dialogic engagement was entirely online suggests that the AWM online adaptation is a potentially viable approach to dialogic teaching and learning. In particular, the relationship between peer argumentation and individual argumentation skills held when the AWM approach concentrated largely on synchronous, electronic dialogue featured in the game phase. Research theory advocates the electronic mode as a medium for facilitating reflection on the externalized thought preserved by the tangible record of exchanges that overtime and with practice progresses to meta-strategic thinking (Hemberger et al., 2017; Iordanou & Rapanta, 2021; Kuhn, 2018a; Kuhn et al., 2016a; Shi et al., 2019). An entirely electronic format may amplify this process and may have contributed to our findings. Also, the fact that practice was spread out over time (e.g., new topics each week), a factor related to enhanced memory (Anderson, 1995) may support knowledge of what strategies to use in particular circumstances. However, we have no way of knowing because we did not examine if participants accessed old dialogues on- or off- workshop hours and if so, for how long. Future work can extend upon our findings by adding measures that inform our understandings of the role of the electronic medium and spacing effects.

Although the relationship between argumentation and argumentive competence held, it was not strong (particularly so for the participants in the comparison condition) and effects of dialogue were at best immediate (particularly for the participants in the experimental condition). Research theory predicts argumentation skill development occurs gradually and remains incomplete after one year in the full, in-person AWM approach (Hemberger et al., 2017; Kuhn et al., 2016a, 2016b; Shi et al., 2019); our findings are accordingly consistent (Mayweg-Paus et al.,

2015; Papathomas & Kuhn, 2017). Research has only begun investigating ways to accelerate development and we wondered, given that current intervention was more limited, if and how the context-specific distinctions played a role. For example, is concentrated dialogic interaction with an opposing-side partner in writing equally as potent as balanced verbal and written discourse with same-side and opposing-side peer-pairs, respectively? In the future, researchers may want to examine whether additional practice in entirely electronic argumentive discourse (in addition to the full, in-person, AWM approach) accelerates development.

These context-specific distinctions leave us with lots of questions concerning regulation processes that potentially have exciting implications for the field. Is an all writing dialogic context more powerful than a mix of verbal and written dialogues? Would results have been different if partners engaged in one-to-one discourse balanced across opposing-side, as well as same-side, partners? Several studies have investigated peer regulation in AWM and found greater gains when students engaged in dialogue with opposing-side peer-pairs, but thinking still expanded when engaging in discourse with same-side peer-pairs (Iordanou & Kuhn, 2020; Kuhn et al., 2019). Additionally, Zillmer and Kuhn (2018) found gains in peer co-regulation when both peers were consistently partnered with their same-side peer. Still, there is much to be learned about regulation processes in dialogic argumentation contexts in one-to-one writing contexts. Future studies could conduct microgenetic studies using activity theory as an appropriate methodological framework for capturing the moment-by-moment dialogic interactions in-situ.

To that end and to our knowledge, this is the first study that adapted the AWM approach into an all-online format, which makes it a promising avenue for exploring the benefits of the online adaptation. Future investigations with the online adaptation have the potential to contribute to what little we know of its unique features (e.g., anonymity). Future studies could

also test and explore additional aspects and features of the pre-game and end-game in the all-online adaptation. Iordanou (2013) devised a hybrid "showdown" during the endgame, which could probably be adapted into an all-online format. Would it be too logistically challenging for partners to work in same-side peer-pairs as well as opposing-side peer-pairs in an all online format? And if so, would one-to-one dialogic interaction with same-side and opposite-side partners be a viable substitute? Future investigations could contribute to a rapidly growing literature on computer-supported-collaborative-learning and furthermore, have the potential to serve as viable alternatives to dialogic education, particularly, in times of crisis.

Finally, our study coding scheme extends upon previous dialectical coding schemes (Mayweg-Paus et al., 2015; Papthomas & Kuhn, 2017), making an important methodological distinction that has implications for evaluating change in counterargument strategy use and argumentive skill development. The key indicator of argumentation skill employed is the use of counterargument as an argumentation strategy and more importantly, the use of stronger argumentive moves that imply developmental change. Counterarguments constitute the core of argument: they strengthen or weaken the force of an argument (Crowell & Kuhn, 2014; Goldstein et al., 2009; Kuhn et al., 2016b; Mayweg-Paus et al., 2015; Papathomas & Kuhn, 2017). They also provide insight into the underlying structure of the argumentive discourse. Previously, Mayweg-Paus et al. (2015) and Papathomas and Kuhn (2017) identified three counterargument strategies in peer dialogues, counter-alternative, counter-critique, and counter-undermine. Mayweg-Paus et al. (2015) identified counter-undermine as the strongest of the three counterarguments; Papathomas and Kuhn (2017) identified counter-critique and counter-undermine as the stronger two of the three counterarguments. We added a fourth strategy, counter-disagree to balance the assessment of strength of argumentive strategy use,

distinguishing the two weaker strategies (counter-disagree and counter-alternative) from the two stronger strategies (counter-critique and counter-undermine) (see Table 14).

Table 14

Counterargument Strategies in Order of Power

	Argument Strategy
Strategy strength (Weak → Strong)	*Counter-Disagree*
	Counter-Alternative
	Counter-Critique
	Counter-Undermine

Use of powerful strategies reflect development in production skills central to argumentive discourse (Crowell & Kuhn, 2014; Goldstein et al., 2009). Novice arguers initially use weak argumentive strategies frequently and seldom use strong strategies (if at all), suggesting that they have yet to realize the structure of argumentive thinking and its productive purposes (Kuhn et al., 2016; Mayweg-Paus et al., 2015; Papathomas & Kuhn, 2017). Extended engagement in dialogic arguing activities supports use of powerful strategies, typically beginning with counter-critique and advancing to counter-undermine, a particularly strong strategy that addresses deeper grounds of disagreement (Crowell & Kuhn, 2014; Goldstein et al., 2009; Mayweg-Paus et al., 2015). Although increasingly frequent use of powerful strategies develops gradually with time and practice, findings from current and earlier research suggest that dialogic interaction with an expert has an immediate and direct effect on production skills (Mayweg-Paus et al., 2015; Papathomas & Kuhn, 2017).

Educational Implications

Often the goal of dialogic argumentation-based curricular interventions is to develop students' argumentive reasoning skills and competencies through collaborative dialogues or arguing activities, ideally to enhance students' abilities to participate in the activities of arguing in more competent ways. Though not exclusive, a main goal is to develop students' argumentive prose via writing. Cognitive and developmental researchers say that developing these competencies are important for learning how to think critically.

Given that younger teenagers spend the majority of their time in school, it is worth exploring the kinds of educational experiences they obtain in school. Experience shapes brain development which, in turn, influences behavior during adolescence, a period of time in which educational experiences become less homogenous as teens begin to invest themselves in managing their own lives (Kuhn, 2006). Findings from research on brain development corroborate the idea that increasing freedom and personal control amplifies teens' experiences and, to an extent, functions to guide or channel behavior (Kuhn, 2006). The adolescent brain becomes more specialized as a result of synaptic pruning that occurs from experience, and self-regulation improves (Giedd et al., 2012). We believe that researchers, educational administrators, and policy-makers alike ought to consider the kinds of educational experiences the adolescent brain has.

Because argumentation programs are often student-centered, aim to develop critical thinking, and focus on the interpersonal interactions amongst peers, they potentially support heterogenous groups of students, as opposed to homogenous groups in traditional learning environments. The dynamic, interactive, and socially situated nature of argumentation programs support variety in students' competencies to develop within their own zones of proximal

development. In contrast to a traditional classroom setting, variety is viewed as a nuisance because those with more advanced competencies will likely be bored and therefore disengaged, and those with underdeveloped competencies will struggle and also not be motivated to persist or engage. Only those whose development is ideally "proximal" will reap the benefits.

Still, research suggests that dialogic learning environments that are student-centered and inquiry-based issue an important caveat: their effectiveness depends upon the readiness and quality of the teachers implementing them (Furtak et al., 2012; Minner et al., 2009). Reform initiatives require teachers to reframe classroom life, shifting their professional roles from traditional 'transmitters' of knowledge to an innovative 'learning coach', but little is known about how teachers learn to promote effective student dialogue and argument that leads to learning (Osborn et al., 2013; Simon et al., 2006). Research on teacher professional development (PD) sheds some insight into the matter. Studies show that teachers hesitate to press for questions or use questions effectively and/or plan for development of extended inquiry when faced with the complexities and challenges of re-orienting classroom teaching practices (McNeill & Knight, 2013; Sampson & Blanchard, 2012). Support for teachers is likely ineffective and/or minimal, which may be one reasons why dialogic approaches are not often employed in formal classroom contexts.

In dialogic teaching and learning, the criteria for achievement is not explicitly "realized or idealized" (Michaels et al., 2008; Resnick et al., 2015; Resnick et al., 2018). The thinker evaluates, synthesizes, and interprets information to solve a problem (Hyytinen et al., 2018) and inherently brings her/his relevant subjective realities, knowledge, and experiences that are often overlooked in traditional classroom settings. Thinkers are given opportunities to coordinate their own subjective realities with objective properties of knowledge. Moreover, thinking is not an

isolated endeavor. Rather it is an open exchange of ideas whereby students hold each other accountable to norms and expectations of behavior that emerge and evolve. People build knowledge together and in so doing develop together (Michaels et al., 2008). Dialogic argumentation provides a pathway for young adolescents to practice and develop higher-order cognitive skills through social engagement with peers and teachers. Through dialogic education, the goals of education promulgated by current reform efforts (e.g., Common Core State Standards, 2010) can be achieved: teach students *how* to think, not *what* to think.

References

Ainsworth, S., Gelmini-Hornsby, G., Threapleton, K., Crook, C., O'Malley, C., & Buda, M. (2011). Anonymity in classroom voting and debating. *Learning and Instruction, 21*(3), 365-378. https://doi.org/10.1016/j.learninstruc.2010.05.001

Alexander, R. J. (2006). *Towards dialogic teaching: Rethinking classroom talk* (3rd ed.). Cambridge: Dialogos.

Alexander, R. J. (2018). Developing dialogic teaching. Genesis, process, trial. *Research Papers in Education, 33*(5), 561-598. https://doi.org/10.1080/02671522.2018.1481140

Anderson, J. R. (1995). *Learning and Memory: An integrated approach.* New York: Wiley.

Asterhan, C. S. C. (2018). Exploring enablers and inhibitors of productive peer argumentation: The role of individual achievement goals and of gender. *Contemporary Educational Psychology*, *54*(1), 66-78. https://doi.org/10.1016/j.cedpsych.2018.05.002

Asterhan, C. S. C., & Babichenko, M. (2015). The social dimension of learning through argumentation: Effects of human presence and discourse style. *Journal of Educational Psychology*, *107*(3), 740–755. https://doi.org/10.1037/edu0000014

Asterhan, C. S. C., & Schwarz, B. (2016). Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist, 51*(2), 164-187. https://doi.org/10.1080/00461520.2016.1155458

Billig, M. (1987). *Arguing and thinking: A rhetorical approach to social psychology.* Cambridge University Press.

Butera, F., & Darnon, C. (2017). Competence assessment, social comparison, and conflict regulation. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (pp. 192-213. The Guildord Press. https://psycnet.apa.org/record/2017-17591-011

Butera, F., & Mugny, G. (1995). Conflict between incompetences and influence of a low-expertise source in hypothesis testing. *European Journal of Social Psychology*, *25*(4), 457–462. https://doi.org/10.1002/ejsp.2420250408

Butera, F., Sommet, N., & Darnon, C. (2019). Sociocognitive conflict regulation: How to make sense of diverging ideas. *Current Directions in Psychological Science*, *28*(2), 145–151. https://doi.org/10.1177/0963721418813986

Chinn, C. A., & Buckland, L. A. (2012). Model-based instruction: Fostering change in evolutionary conceptions and in epistemic practices. In K. S. Rosengren, E. M., Evans, S. Brem, & G. M. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in*

teaching and learning about evolution (pp. 211-232). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199730421.003.0010

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Common Core State Standards Initiative. (2010). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*. http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Crowell, A., & Kuhn, D. (2014). Developing dialogic argumentation skills: A 3-year intervention study. *Journal of Cognition and Development, 15*(2), 363-381. https://doi.org/10.1080/15248372.2012.725187

Darnon, C., Doll, S., & Butera, F. (2007). Dealing with a disagreeing partner: Relational and epistemic conflict elaboration. *European Journal of Psychology of Education, 22*(3), 227–242. https://doi.org/10.1007/BF03173423

Felton, M. (2004). The development of discourse strategies in adolescent argumentation. *Cognitive Development, 19*(1), 35-52. https://doi.org/10.1016/j.cogdev.2003.09.001

Felton, M., & Kuhn, D. (2001). The development of argumentive discourse skill. *Discourse Processes, 32*(2&3), 135-153.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research, 82*(3), 300-329. https://doi.org/10.3102/0034654312457206

Giedd, J. N., Stockman, M., Weddle, C., Liverpool, M., Wallace, G. L., Lee, N. R., Lalonde, F., & Lenroot, R. K. (2012). Anatomic magnetic resonance imaging of the developing child and adolescent brain. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The adolescent brain: Learning, reasoning, and decision making.* (pp. 15–35). American Psychological Association. https://doi.org/10.1037/13493-001

Goldstein, M., Crowell, A., & Kuhn, D. (2009). What constitutes skilled augmentation and how does it develop? *Informal Logic, 29*(4), 379-395.

Graff, G. (2003). *Clueless in academe: How schooling obscures the life of the mind.* Yale University Press.

Hemberger, L., Kuhn, D., Matos, F., & Shi, Y. (2017). A dialogic path to evidence-based argumentive writing. *Journal of the Learning Sciences, 26*(4), 575-607. https://doi.org/10.1080/10508406.2017.1336714

Hilbe, J. M. (2011a). Negative binomial regression. *Negative Binomial Regression* (2nd ed.). Cambridge University Press.

Hilbe, J. M. (2011b). Overdispersion. *Negative Binomial Regression* (2nd ed.). Cambridge University Press.

Hilbe, J. M. (2011c). Poisson regression. *Negative Binomial Regression* (2nd ed.). Cambridge University Press.

Hyytinen, H., Toom, A., & Postareff, L. (2018). Unraveling the complex relationship in critical thinking, approaches to learning and self-efficacy beliefs among first-year educational science students. *Learning and Individual Differences, 67,* 132-142. https://doi.org/10.1016/j.lindif.2018.08.004

Iordanou, K. (2013). Developing face-to-face argumentation skills: Does arguing on the computer help? *Journal of Cognition and Development, 14*(2), 292-320. https://doi.org/10.1080/15248372.2012.668732

Iordanou, K., & Kuhn, D. (2020). Contemplating the opposition: Does a personal touch matter? *Discourse Processes, 57*(4), 343-359. https://doi.org/10.1080/0163853X.2019.1701918

Iordanou, K., & Rapanta, C. (2021). "Argue with me": A method for developing argument skills. *Frontiers in Psychology.* https://doi.org/10.3389/fpsyg.2021.631203

Iordanou, K., Kuhn, D., Matos, F., Shi, Y., & Hemberger, L. (2019). Learning by arguing. *Learning and Instruction, 63,* 101-107. https://doi.org/10.1016/j.learninstruc.2019.05.004

Jordan-O'Reilly, M. (2013). *Social comparison: How does it work?(Writings in psychology book 3).*

Kolikant, Y., & Pollack, S. (2015). The dynamics of non-convergent learning with a conflicting other: Internally persuasive discourse as a framework for articulating successful collaborative learning. *Cognition and Instruction, 33*(4), 322–356. https://doi.org/10.1080/07370008.2015.1092972

Kozulin, A. (1986). The concept of activity in Society Psychology: Vygotsky, his disciples and critics. *American Psychologist, 41*(3), 264-274. https://doi.org/10.1037/0003-066X.41.3.264

Križan, Z., & Gibbons, F. X. (Eds). (2014). *Communal functions of social comparison.* Cambridge University Press.

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

Kuhn, D. (2006). Do cognitive changes accompany developments in the adolescent brain? *Perspectives on Psychological Science, 1*(1), 59–67. https://doi.org/10.1111/j.1745-6924.2006.t01-2-.x

Kuhn, D. (2018a). A role for reasoning in a dialogic approach to critical thinking. *Topoi, 37*(1), 121-128. https://doi.org/10.1007/s11245-016-9373-4

Kuhn, D. (2018b). *Building our best future: Thinking critically about ourselves and our world* (2nd ed). Wessex Press Publishing Co.

Kuhn, D. (2019). Critical thinking as discourse. *Human Development, 62*(3), 146-164. https://doi.org/10.1159/000500171

Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychological Science, 22*(4), 545-552. https://doi.org/10.1177/0956797611402512

Kuhn, D., & Modrek, A. (2021). Mere exposure to dialogic framing enriches argumentive thinking. *Applied Cognitive Psychology, 35*(5), 1349-1355. https://doi.org/10.1002/acp.3862

Kuhn, D., & Moore, W. (2015). Argumentation as core curriculum. *Learning: Research and Practice, 1*(1), 1-13. https://doi.org/10.1080/23735082.2015.994254

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

Kuhn, D., & Zillmer, N. (2015). Developing norms of discourse. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 77-86). American Educational Research Association. www.jstor.org/stable/j.ctt1s474m1

Kuhn, D., Feliciano, N., & Kostikinia, D. (2019). Engaging contemporary issues as practice for citizenship. *The Social Studies, 110*(5), 207-219. https://doi.org/10.1080/00377996.2019.1625856

Kuhn, D., Goh, W., Iordanou, K., & Shaenfield, D. (2008). Arguing on the computer: A microgenetic study of developing argumentation skills in a computer-supported environment. *Child Development, 79,* 1310-1328. https://doi.org/10.1111/j.1467-8624.2008.01190.x

Kuhn, D., Hemberger, L., & Khait, V. (2016a). *Argue with me: Argument as a path to developing students' thinking and writing* (2nd ed.). Routledge.

Kuhn, D., Hemberger, L., & Khait, V. (2016b). Tracing the development of argumentive writing in a discourse-rich context. *Written Communication, 33*(1), 92-121. https://doi.org/10.1177/0741088315617157

Kuhn, D., Wang, Y., & Li, H. (2011). Why argue? Developing understanding of the purpose and value of argumentive discourse. *Discourse Processes, 48*(1), 26-49. https://doi.org/10.1080/01638531003653344

Kuhn, D., Zillmer, N., Crowell, A., & Zavala, J. (2013). Developing norms of argumentation: Metacognitive, epistemological, and social dimensions of developing argumentive competence. *Cognition and Instruction, 31*(4), 456-496. https://doi.org/10.1080/07370008.2013.830618

Littleton, K., & Mercer, N. (2013). *Interthinking: Putting talk to work* (1st ed.). Routledge. https://doi.org/10.4324/9780203809433

Lombardi, D., Nussbaum, E. M., & Sinatra, G. M. (2016). Plausibility judgements in conceptual change and epistemic cognition. *Educational Psychologist, 51*(1), 35-56. https://doi.org/10.1080/00461520.2015.1113134

Lombardi, D., Sinatra, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction, 27,* 50-62. https://doi.org/10.1016/j.learninstruc.2013.03.001

Lun, J., Sinclair, S., Whitchurch, E. R., & Glenn, C. (2007). (Why) do I think what you think? Epistemic social tuning and implicit prejudice. *Journal of Personality and Social Psychology*, *93*(6), 957–972. https://doi.org/10.1037/0022-3514.93.6.957

Macagno, F. (2016). Argument relevance and structure. Assessing and developing students' uses of evidence. *International Journal of Educational Research, 79,* 180-194. https://doi.org/10.1016/j.ijer.2016.07.002

Macagno, F., Maweg-Paus, E., & Kuhn, D. (2014). Argumentation theory in education studies: Coding and improving students' argumentative strategies. *Topoi, 34,* 523-537. https://doi.org/10.1007/s11245-014-9271-6

Matos, F. (2021). Collaborative writing as a bridge from peer discourse to individual argumentative writing. *Reading and Writing, 34,* 1321-1342. https://doi.org/10.1007/s11145-020-10117-2

Maweg-Paus, E., Macagno, F. & Kuhn, D. (2015). Developing argumentation strategies in electronic dialogs: Is modeling effective? *Discourse Processes, 53*(4), 280-297. https://doi.org/10.1080/0163853X.2015.1040323

McNeill, K. L., & Knight, A. M. (2013). Teachers' pedagogical content knowledge of scientific argumentation: The impact of professional development on K-12 teachers. *Science Teacher Education, 97*(6), 936-972. https://doi.org/10.1002/sce.21081

Mercer, N. (1995). *The guided construction of knowledge: Talk amongst teachers and learners.* Multilingual Matters.

Mercer, N., & Howe, C. (2012). Explaining the dialogic processes of teaching and learning: The value and potential of sociocultural theory. *Learning, Culture and Social Interaction, 1*(1), 12-21. https://doi.org/10.1016/j.lcsi.2012.03.001

Mercer, N., & Littleton, K. (2007). *Dialogue and development of children's thinking. A sociocultural approach* (1st ed.). Routledge. https://doi.org/10.4324/9780203946657

Mercer, N., Wegerif, R., & Major, L. (Eds.). (2020). *The Routledge international handbook of research on dialogic education* (1st ed.). Routledge. https://www.routledge.com/The-Routledge-International-Handbook-of-Research-on-Dialogic-Education/Mercer-Wegerif-Major/p/book/9781138338517

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and civic life. *Studies in Philosophy and Education, 27*(4), 283-297. https://doi.org/10.1007/s11217-007-9071-1

Minner, D. D., Levy, A. J., & Curry, J. (2009). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984-2002. *Journal of Research in Science Teaching, 47*(4), 474-496. https://doi.org/10.1002/tea.20347

Murphy, P. K., Greene, J. A., Firetto, C. M., Hendrick, B. D., Li, M., Montalbano, C., & Wei, L. (2018). Quality talk: Developing students' discourse to promote high-level comprehension. *American Educational Research Association, 55*(5), 1113-1160. https://doi-org.tc.idm.oclc.org/10.3102/0002831218771303

National Association of Independent Schools. (n. d.). *NAIS Demographic Data* (Market View) [Data and Analysis for School Leadership (DASL)]. https://www.nais.org/analyze/data-and-analysis-for-school-leadership-(dasl)/demographic-center/

National Center for Education Statistics. (2011). *The nation's report card, Writing 2011:National Assessment of Educational Progress at grades 8 and 12* (NCES 2012-470). U.S. Department of Education & Institute of Education Sciences. https://www.nationsreportcard.gov/writing_2011/

National Center for Education Statistics. (2019 – 2020). *Common Core of Data* (CCD) [Unpublished raw data]. CCD Public School Data 2019-2020 School Year. https://nces.ed.gov/ccd/schoolsearch/

NCSS Statistical Software. (2021). Negative binomial Regression. https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf

Nussbaum, E. M. (2008). Using argumentation vee diagrams (AVDs) for promoting argument-counterargument integration in reflective writing. *Journal of Educational Psychology, 100*(3), 549-565. https://doi.org/10.1037/0022-0663.100.3.549

Nussbaum, E. M. (2021). Critical integrative argumentation: Toward complexity in students' thinking [Presidential Address]. *Educational Psychologist, 56*(1), 1-17. https://doi.org/10.1080/00461520.2020.1845173

Nussbaum, E. M., & Putney, L. G. (2020). Learning to use benefit-cost arguments: A microgenetic study of argument-counterargument integration in an undergraduate seminar course. *Journal of Educational Psychology, 112*(3), 444-465. https://doi.org/10.1037/edu0000412

Nussbaum, E. M., & Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *The Journal of Experimental Education, 76*(1), 59-92. https://doi.org/10.3200/JEXE.76.1.59-92

Nussbaum, E. M., Dove, I. J., Slife, N., Kardash, C. M., Turgut, R., & Vallett, D. (2019). Using critical questions to evaluate written and oral arguments in an undergraduate general education seminar: A quasi-experimental study. *Reading and Writing, 32,* 1531-1552. https://doi.org/10.1007/s11145-018-9848-3

Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C., & Richardson, K. (2013). Learning to argue: A study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching, 50*(3), 315-347. https://doi.org/10.1002/tea.21073

Papathomas, L., & Kuhn, D. (2017). Learning to argue via apprenticeship. *Journal of Experimental Child Psychology, 159,* 129-139. https://doi.org/10.1016/j.jecp.2017.01.013

Piaget, J. (1952). *The origins of intelligence in children.* (M. Cook, Trans.). W. W. Norton & Co. https://doi.org/10.1037/11494-000

Purcell, K., Buchanan, J., & Friedrich, L. (2013, July 16). *The impact of digital tools on student writing and how writing is taught in schools*. Pew Research Center. https://www.pewresearch.org/internet/2013/07/16/the-impact-of-digital-tools-on-student-writing-and-how-writing-is-taught-in-schools/

Quiamzade, A., Mugny, G., Cléopas, A. D., & Buchs, C. (2003). Interaction styles and expert social influence. *European Journal of Psychology of Education, 18,* 389–404. https://doi.org/10.1007/BF03173243

Resnick, L. B., Asterhan, C. S. C., & Clarke, S. N. (Eds.). (2015). *Socializing intelligence through academic talk and dialogue.* American Educational Research Association. www.jstor.org/stable/j.ctt1s474m1

Resnick, L., Asterhan, C. S. C., Clarke, S. N., & Schantz, F. (2018). Next generation research in dialogic learning. In G. E. Hall, D. M. Gollnick, & L. F. Quinn (Eds.), *Handbook of Teaching and Learning* (pp. 323-338). Wiley-Blackwell.

Reznitskaya, A., & Wilkinson, I. A. G. (2017). Truth matters: Teaching young students to search for the most reasonable answer. *Phi Delta Kappan, 99*(4), 33-38. https://doi.org/10.1177/0031721717745550

Reznitskaya, A., Glina, M., Carolan, B., Michaud, O., Rogers, J., & Sequeira, L. (2012). Examining transfer effects from dialogic discussions to new tasks and contexts. *Contemporary Educational Psychology, 37*(4), 288-306. https://doi.org/10.1016/j.cedpsych.2012.02.003

Reznitskaya, A., Kuo, L., Clark, A., Miller, B., Jadallah, M., Anderson, R. C., & Nguyen-Jahiel, K. (2009). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education, 39*(1), 29-48. https://doi.org/10.1080/03057640802701952

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

Scribner, S., & Cole, M. (1981). *The psychology of literacy.* Cambridge, MA: Harvard University Press.

Shi, Y. (2019). Enhancing evidence-based argumentation in a mainland China middle school. *Contemporary Educational Psychology, 59*(11), 1-18. https://doi.org/10.1016/j.cedpsych.2019.101809

Shi, Y., Matos, F., & Kuhn, D. (2019). Dialog as a bridge to argumentive writing. *Journal of Writing Research, 11*(1), 107-129. https://doi.org/10.17239/jowr-2019.11.01.04

Simon, S., Erduran, S., & Osborne, J. (2006). Learning to teach argumentation: Research and development in the science classroom. *International Journal of Sceince Education, 28*(2), 235-260. https://doi.org/ 10.1080/09500690500336957

Van Eemeren, F., & Grootendorst, R. (1992). *Argumentation, communication and fallacies: A pragma-dialectical perspective* (1st ed.). Routledge.

Ver Hoef, J. M., & Boveng, P. K. (2007). Quasi-Poisson vs. Negative Binomial Regression: How should we model overdispersed count data? *Ecology, 88*(11), 2766-2772. http://www.jstor.org/stable/27651434

Vygotsky, L. S. (1978). *Mind in Society: The development of higher psychological processes*. (M. Cole, V. John-Steiner, S., Scribner, E. Souberman, Eds). Harvard University Press.

Vygotsky, L. S. (1987). *The collected works of L. S. Vygotsky: Problems of general psychology* (R. W. Rieber & A. S. Carton, Eds.). Plenum Press. (Original work published in 1937)

Walton, D. (1989). Dialogue theory for critical thinking. *Argumentation, 3*, 169-184. https://doi.org/10.1007/BF00128147

Walton, D. (2005). How to evaluate argumentation using schemes, diagrams, critical questions and dialogues. *Studies in Communication Sciences*, 51-74. http://doi.org/10.5169/seals-790943

Walton, D. (2014). *Dialog theory for critical argumentation.* John Benjamins.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes.* Cambridge University Press.

Zillmer, N., & Kuhn, D. (2018). Do similar-ability peers regulate one another in a collaborative discourse activity? *Cognitive Development, 45,* 68-76. https://doi.org/10.1016/j.cogdev.2017.12.002

Appendix

Supplementary Data Demonstrating Overdispersion

As demonstrated in tables A1 and A2 below, the data for both animal research and transfer topic essays are non-normal (positive skewness) and are overdispersed (outcome variable variances are approximately two to three times greater than the corresponding outcome variable means). (Note: Evidence-based units were re-counted based on whether the functional use of the unit did/did not include evidence).

Table A1

Animal Research Essay Descriptives by Group

Condition	Outcome Variable	Mean	n	Median	Min.	Max.	SD	Skewness	Kurtosis
Experimental	Support-Own	4.81	26	4.50	1	16	3.45	1.35	2.98
	Weaken-Own	0.65	26	0.00	0	4	1.02	2.01	4.28
	Support-Other	0.69	26	0.00	0	5	1.23	2.20	5.30
	Weaken-Other	1.96	26	1.50	0	9	2.65	1.98	3.23
	Total Idea Units	8.12	26	7.00	1	25	5.86	1.23	1.46
	Total Evidence	3.04	26	2.50	0	11	2.95	1.39	1.52
	Shared Evidence	1.15	26	1.00	0	4	1.41	1.02	-0.23
Comparison	Support-Own	4.26	31	3.00	0	14	3.54	1.26	1.34
	Weaken-Own	0.35	31	0.00	0	3	0.76	2.27	4.79
	Support-Other	0.29	31	0.00	0	2	0.59	1.96	2.97
	Weaken-Other	2.03	31	1.00	0	8	2.15	1.18	0.93
	Total Idea Units	6.94	31	5.00	0	24	6.07	1.46	1.70
	Total Evidence	2.74	31	2.00	0	21	4.21	2.97	11.41
	Shared Evidence	0.65	31	0.00	0	4	1.14	1.77	2.09

Table A2

Transfer Topic Essay Descriptives by Group

Condition		Outcome Variable	Mean	n	Median	Min.	Max.	SD	Skewness	Kurtosis
	Experimental	Support-Own	2.13	23	2.00	0	6	1.69	1.02	0.12
		Weaken-Own	0.13	23	0.00	0	1	0.34	2.35	3.86
		Support-Other	0.39	23	0.00	0	2	0.66	1.50	1.20
		Weaken-Other	1.70	23	1.00	0	9	2.32	1.88	3.53
		Total Idea Units	4.35	23	3.00	1	16	3.47	1.81	4.57
		Total Evidence	1.17	23	0.00	0	13	2.92	3.49	13.14
		Shared Evidence	—	23	—	—	—	—	—	—
	Comparison	Support-Own	2.07	30	2.00	0	7	1.72	0.89	0.80
		Weaken-Own	0.03	30	0.00	0	1	0.18	5.48	30.00
		Support-Other	0.20	30	0.00	0	2	0.55	2.76	6.73
		Weaken-Other	0.67	30	0.00	0	4	0.99	1.87	3.73
		Total Idea Units	2.97	30	2.00	0	13	2.74	2.00	5.28
		Total Evidence	0.63	30	0.00	0	10	2.03	4.01	17.00
		Shared Evidence	—	30	—	—	—	—	—	—