JOINT REFLECTION PROMOTES STUDENTS’ USE OF EVIDENCE IN ARGUMENTIVE WRITING

Yuchen Shi

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
2018
ABSTRACT

Joint Reflection Promotes Students’ Use of Evidence in Argumentative Writing

Yuchen Shi

A basic component of argumentative writing is the coordination of claims with evidence bearing on them. Deep engagement in dialogic argumentation has been found to facilitate development of beginning students’ individual written argument. Despite progress in several respects following such engagement, in their argumentative writing middle-school writers frequently ignore evidence incongruent with their claims -- a violation of norms of skilled argument.

The present research examines the effectiveness of engaging middle-school student dyads in joint meta-level reflection on the use of evidence in their argumentation, both anticipating its potential use and evaluating its actual use. A total of 54 Chinese 7th graders participated in a dialogic argument curriculum in 33 class sessions over four months. For each of three successive topics, evidence both congruent and incongruent with a dyads’ position on the topic was made available for their use. Half of the participants were assigned to an Evidence Reflection and Argument Practice (ER+AP) condition, in which in addition the dyad was prompted to discuss verbally and jointly complete reflection sheets regarding their evidence use. The other half of participants served in an Argument Practice (AP) condition, identical except for omission of the Evidence Reflection component.

Analysis of participants’ individual written essays on the topic at the end of their engagement with each topic revealed superior performance on the part of the ER+AP group, with the reflection component enhancing their addressing evidence both congruent and incongruent with their claims. However, this happened only slowly. The superiority of the ER+AP group was
most decisive by the last topic, when members of the ER+AP students also demonstrated an
ability to connect two pieces of evidence serving conflicting argumentive functions.

Fifty additional students participated in a control condition, included for the purpose of
comparing their performance to that of the intervention students on a topic new to both groups.
Both the ER+AP and AP intervention groups showed superior performance relative to the
control group in including evidence congruent with their own position in their essays. Only the
ER+AP group, however, showed superiority in addressing evidence incongruent with their
position.

Analysis of responses students provided to the evidence reflection sheets revealed
developmental patterns over time, and explicated the underlying mechanism driving ER+AP
students’ superior performance. Theories regarding the interiroization of cognition from inter- to
intra-mental planes, as well as the supportive effects of meta-level engagement on transfer of
skills, are invoked in accounting for the findings.
# TABLE OF CONTENTS

LIST OF TABLES.............................................................................................................................................. iii
LIST OF FIGURES........................................................................................................................................... iv
ACKNOWLEDGEMENTS................................................................................................................................. v

CHAPTER 1: INTRODUCTION & LITERATURE REVIEW................................................................................. 1

Introduction .................................................................................................................................................... 1
  Argumentation and Learning ....................................................................................................................... 2
    Overview of the field .................................................................................................................................. 2
    Arguing to learn vs. Learning to argue ..................................................................................................... 6
    A dialogue-based argument curriculum ................................................................................................. 8
  Contribution of the present study ........................................................................................................... 11

The challenge of arguing from evidence ................................................................................................... 13
  Distinguish evidence from explanation .................................................................................................... 13
  Self-serving evaluation of evidence ......................................................................................................... 15

Metacognition and metacognitive training ............................................................................................... 18
  Models of metacognition ......................................................................................................................... 18
  Metacognition in developmental and cognitive research ........................................................................... 18
  Metacognitive training in educational research ....................................................................................... 20

Hypotheses of the present study .............................................................................................................. 23

CHAPTER 2: METHODS ............................................................................................................................. 24

Participants ................................................................................................................................................ 24

Design ......................................................................................................................................................... 24

Procedure ................................................................................................................................................... 26
  Pre- & Post-assessment ........................................................................................................................... 27
  Intervention topics .................................................................................................................................... 27
  Intervention sequence of activities ......................................................................................................... 28
  Condition manipulation .......................................................................................................................... 32

CHAPTER 3: RESULTS .................................................................................................................................. 34

Coding of essays ........................................................................................................................................ 35

Analysis of intervention topic essays ....................................................................................................... 41

Analysis of transfer essays ....................................................................................................................... 52
Qualitative portrayal of change in essays..............................................................60
Analysis of evidence reflection sheets.................................................................68
CHAPTER 4: DISCUSSION..................................................................................73
Principal findings.................................................................................................73
Educational and theoretical implications............................................................75
Limitations and future directions........................................................................83
REFERENCES.....................................................................................................85
APPENDIX A.......................................................................................................92
APPENDIX B.......................................................................................................94
APPENDIX C.......................................................................................................96
APPENDIX D......................................................................................................100
LIST OF TABLES

Table 1. Means (and SDs) on school diagnostic test by condition ................................................. 25
Table 2. List of intervention topics ........................................................................................................... 28
Table 3. Coding scheme for categorizing evidence-based units .............................................................. 37
Table 4. Verbatim (translated) examples of functional evidence-based units serving different argumentive goals ........................................................................................................................................... 38
Table 5. Percentage of students who ever made a certain type evidence-based claim by condition and topic .................................................................................................................................................... 43
Table 6. Percentage of students who ever made an evidence-based However units and full evidence-based However units by condition and topic ........................................................................................................ 50
Table 7. Descriptive statistics for measures in pre- and posttests by condition ........................................... 53
Table 8. Percentage of students who ever made a certain type evidence-based claim by condition at pretest and posttest ............................................................................................................................................. 59
Table 9. Percentage of students who ever made an evidence-based However units and full evidence-based However units by condition at pretest and posttest ..................................................................................... 60
Table 10. Percentage analysis for evidence reflection sheet – part B .................................................................... 72
LIST OF FIGURES

Figure 1. Four argumentive functions of reasons and evidence ........................................... 10

Figure 2. Diagrammatic representation of the essay coding scheme ..................................... 36

Figure 3. Diagrammatic representation of however unit ...................................................... 40

Figure 4. Mean number of different types of evidence-based units by condition at Topic 1 ................................................................................................................................. 43

Figure 5. Mean number of different types of evidence-based units by condition at Topic 2 ....................................................................................................................................... 44

Figure 6. Mean number of different types of evidence-based units by condition at Topic 3 ....................................................................................................................................... 46

Figure 7. Mean percentage of idea units served as claim-congruent and claim-incongruent functional evidence-based claims by condition and topic ........................................... 49

Figure 8. Mean number of evidence-based however units and full evidence-based however units by condition and topic .................................................................................................................. 50

Figure 9. Mean number of claim-congruent functional evidence-based units by condition at pretest and posttest .................................................................................................................. 54

Figure 10. Mean number of claim-incongruent functional evidence-based units by condition at pretest and posttest .................................................................................................................. 55

Figure 11. Mean number of evidence-based however units by condition at pretest and posttest ....................................................................................................................................... 57

Figure 12. Mean number of full evidence-based however units by condition at pretest and posttest ....................................................................................................................................... 58
ACKNOWLEDGEMENTS

I would like to extend my first and foremost appreciation to my beloved advisor Deanna Kuhn, who took me on as her doctoral student several years ago and guided me with unparalleled professionalism and compassion. Without her, I would not have navigated so well all the ups and downs along this PhD journey. I also want to thank two of my dearest lab members – Flora Matos and Laurie Hemberger. They supported me like family during the past four years. They believed in me and were always there whenever I needed a pick-me-up.

I also would like to thank the teachers and students at the school in which the study was conducted. The work was made possible only with the remarkable support from the school administration and the dedicated work of the participating students and teachers.

Lastly, I want to thank my parents for bringing me into this world, for teaching me and giving me the space to think for myself, for supporting me when I wanted to chase my dreams, and for always surrounding me with unconditional and boundless love.
Introduction

Argumentive writing poses a challenge to students of all ages, with most evidence indicating that skilled argumentive writing is not readily achieved by the majority of students even at the college level (Graham & Perin, 2007; Wolfe, 2011). Among others, two weaknesses have been identified in young adolescents’ argumentive writing. The first is their struggle with constructing two-sided arguments as they typically concentrate their attention on exposition of their own claims, essentially ignoring the opponent’s claims (Felton & Kuhn, 2001; Kuhn & Udell, 2003; Kuhn et al., 2008). In fact, in individual reasoning, this tendency to construct one-sided arguments might have an evolutionary root as solitary reasoners might be inherently lazy, biased, and content with generic, superficial reasons (Mercier & Sperber, 2011; Mercier, Boudry, Paglieri, & Trouche, 2017).

Another challenge young adolescents encounter is to distinguish evidence and explanation in support of their claims (Brem & Rips, 2000; Kuhn, 1991), a challenge that sometimes manifests in young adolescents’ difficulty with drawing on relevant evidence to inform their claims (Kuhn & Moore, 2015). Research on individuals’ coordination of prior beliefs and new evidence (Stanovich & West, 1997), as well as on the self-serving influence of adolescents’ evaluation of belief relevant evidence (Klacznyski & Gordon, 1996), has demonstrated the effects of my-side bias on individuals’ processing of new evidence, with individuals subjecting belief-inconsistent evidence to more rigorous processing, while accepting at face value belief-consistent evidence. Previous research also indicates that even following extended engagement in the argument curriculum, students rarely acknowledge or address evidence incongruent with their claims (Hemberger et al., 2017; Kuhn & Moore, 2015).
Students fail to address claim-incongruent evidence because they do not realize that refuting it can improve the persuasiveness of their arguments (O’Keefer, 1999). An alternative explanation is that students lack the skills to refute such evidence and are thus reluctant to acknowledge it. However, successful argumentative writing requires addressing and weighing arguments and evidence both congruent and incongruent with one’s own claims, so that a comprehensive conclusion based on such two-sided evaluation can be reached (Nussbaum & Kardash, 2005; Nussbaum & Edwards, 2011).

In the present study, we advocate a dialogue-based argument curriculum with the goal of improving Chinese middle-school students’ skills in addressing evidence both congruent and incongruent with their claim in argumentative writing. We argue that combining extended practice in evidence use during argumentative dialogue with joint reflection on evidence use constitutes the most optimal path. In the following section, we review research on argumentation and learning, with particular focus on our dialogue-based curriculum and contribution of the present study to the curriculum.

**Argumentation and learning**

*Overview of the field*

Argumentation is a verbal and social activity of reason aimed at increasing (or decreasing) the acceptability of a controversial standpoint for the listener or reader, by putting forward a constellation of propositions intended to justify (or refute) the standpoint before a rational judge. (p. 5)

Defined by van Eemeren et al. (1996) as above, argumentation as an intellectual and social activity has captured the scholarly interests of Western culture for the past thousands of years. The cultural legacy of interest in argumentation can be dated as far back as to Socrates’ dialogue and Aristotle’s advancements in logic, dialectic, and rhetoric. In the contemporary
world, argumentation theorists such as Walton (1989), van Eemeren and Grootendorst (1992) continued to make remarkable strides in the studies of dialogue and argumentation from a philosophical standpoint.

In the last three decades or so, psychologists and educational researchers have increasingly recognized argumentation as one of the fundamental intellectual skills schools and colleges should seek to promote. Cognitive scientists have in fact claimed that arguing is not just central to human thinking and reasoning but its central objective (Mercier & Sperber, 2011). As a result, an increasing number of psycho-educational scholars have emphasized the importance of designing curriculum activities that promote quality argumentation in classrooms. The interest in argumentation in classrooms has since flourished into a prominent field encompassing theoretical analyses, empirical research, and policy recommendations. The rationale behind this vibrant focus on the educational value of argumentation is that embedding argumentation into school curriculum improves students’ knowledge and deep understanding of subject content (Asterhan & Schwarz, 2016), their collaborative problem solving skills (Chinn & Clark, 2013), their reasoning and critical thinking skills (Halpern, 1998; Kuhn, 2016a), their expository writing skills (Kuhn & Crowell, 2011), and their civic engagement in a democratic society (Kuhn, 2005).

Moreover, high-profile standards including the 2010 Common Core Standards initiatives and Next Generation Science Standards have all featured skilled argumentation as a prominent educational goal in the 21st century. While the so-called four Cs - critical thinking, communication, collaboration, and creativity - have been stipulated as key 21st–century skills (National Education Association, 2014), researchers have also recognized learning, or inquiry learning, as a key 21st-Century skill (Kuhn, 2016b). Inquiry learning entails engagement in self-regulated discovery of new knowledge not identified in advance, as opposed to asking learners to
passively reproduce information previously transmitted to them, as traditional models of learning depict. Educators promoting inquiry learning in classrooms have often sought to build argumentation into their inquiry-based classroom interventions (Chinn & Clark, 2013). Argumentation in collaborative inquiry-based settings allows learners to transact on each other’s ideas through a process of social transmission, followed by internalization from inter-personal to intra-personal planes, as the sociocultural model posits (Vygotsky, 1978; Rogoff, 1990). It also helps to bring to the surface differences in understanding and encourages learners’ to explore these alternative views, leading to learning gains, as the sociocognitive conflict model predicts (Ames & Murray, 1982; Doise, Mugny, & Perret-Clermont, 1975).

Inquiry and argumentation can be situated within broader pedagogical approaches of the last three decades that emphasize productive classroom discourse in general (Resnick, Asterhan, & Clark, 2015; Mercer & Littleton, 2007), and dialogic learning in particular (Resnick & Schantz, 2015; Mercer, Hennessy, & Warwick, 2017). Productive discourse-based learning requires careful orchestration of talk and tasks in academic learning (Michaels, O’Connor, & Resnick, 2008), and it encourages productive student-to-student, and teacher-to-student discussions that emphasize participation, accountability, and reasoning, with the goal of developing students’ use of language as a psychological tool for learning and problem solving (Mercer et al., 2017). Resnick and Schantz (2015) summarize the dialogic learning approach as a sustained teacher-led but student-owned process of shared reasoning, which ultimately leads to more fully-developed, evidence-supported reasoning.

Productive student-centered classroom discourse that removes teachers from the center of discussion has come under various labels, such as “Collaborative Reasoning” (e.g., Anderson, Chinn, Waggoner, & Nguyen, 1998), “Accountable Talk” (Resnick, Michaels, & O’Connor,
Researchers who have followed this dialogic lead have most frequently focused their investigations on whole-classroom or small-group student discourse.

The work presented here, in contrast, emphasizes dyadic peer discourse (i.e., one dyad arguing against an opposing-side dyad). It follows the discourse-based dialogic approach in classroom research, however, and it highlights the educational value of dispositive exchange of arguments and evidence. In recent years, less adversarial versions of argumentation, such as deliberative argumentation in which students criticize each other’s explanations without the goal of Persuading others, have also been proposed and tested empirically, with mixed results regarding effectiveness compared to the more confrontational, dispositive type of argumentation promoted in the present study (Garica-Mila, Gilabert, Erduran, & Felton, 2013; Iordanou & Constantinou, 2015).

In our line of research, while not denying the potential of deliberative argumentation, we promote argumentation with the discourse goal of deciding between opposing positions. It has the benefit of roots in early development. Research has indicated that producing justifications, counterarguments, and rebutting counterarguments are all present in the discourse of young children (Anderson, Chinn, Chang, Waggoner, & Yi, 1997). Moreover, engaging in controversy can be intrinsically motivating (Ryan & Deci, 2000) to students and potentially leads to ‘epistemic curiosity’ (Berlyne, 1954). On the other hand, research has indicated that collaborative discourse, upon which deliberative argumentation is based, does not come naturally and may be effective only following some delay (Howe, McWilliam, & Cross, 2005), in part because it does not require participants to directly address and question one another’s statements.
Thus, we view students’ engagement in extended oppositional argumentation as opening the
door to other types of discourse with greater emphasis on peer collaboration.

Arguing to learn versus learning to argue

Psycho-educational research on argumentation has tended to fall into two strands,
learning to argue and arguing to learn (Asterhan & Schwarz, 2016). Researchers adopting the
arguing to learn stance mostly come from disciplinary education backgrounds and view
argumentation as a tool to improve students’ domain-specific content learning, such as science
learning (Duschl & Osborne, 2002; Sandoval & Millwood, 2005; Zohar & Nemet, 2002;
Danielson, Sinatra, & Kendeou, 2016), history learning (Wiley & Voss, 1999; De La Paz, 2005),
and mathematics learning (Schwarz, Hershkowitz, & Prusak, 2010). However, Asterhan and
Schwartz (2016) rightfully pointed out that despite widespread interest and abundant research,
evidence of a causal link between argumentation and domain-specific knowledge gains is still scant.

In contrast, researchers who take the learning to argue approach advocate the importance
and feasibility to develop students’ argumentation skills, with focus on teaching the components
of argumentation and effectively engaging students in the practice of argumentation through
extended intervention programs (Dong, Anderson, Kim, & Li, 2008; Kuhn & Crowell, 2011;
Kuhn, Hemberger, & Khait, 2016b; Nussbaum & Schraw, 2007). Within this learning to argue
approach, there is further variation in the level of direct support provided to students. At one end
of the continuum, emphasis is placed on direct instruction and teachers are expected to provide
explicit teaching and scaffolding on argument skills (Larson, Britt, & Kurby, 2009; Hefter et al.,
2017); at the other end of the continuum, emphasis is placed on minimizing instructor’s role
while students engage in extended argumentive discourse with peers (Crowell & Kuhn, 2014). Lying in between these two ends include interventions that provide a limited degree of teacher scaffolding, such as those that promote learning through apprenticeship by engaging students in extended argumentive discourse with an expert adult arguer (Papathomas & Kuhn, 2017; Mayweg-Paus, Macagno & Kuhn, 2015). The variation in the level of direct instruction provided to students within the learning to argue approach reflects the larger debate regarding the effectiveness of direct instruction versus discovery learning (Klahr & Nigam, 2004; Dean & Kuhn, 2006).

The divide between the arguing to learn and learning to argue approaches, with the former focusing on acquisition of content knowledge and the latter on development of inquiry strategies and ‘thinking dispositions’ (Stanovich & West, 1997), reflects “the fundamental dilemma of education,” as Richard Elmore has stated, which is “how to introduce learners to complex bodies of knowledge and expertise while at the same time placing them in the position of assuming responsibility and control over the process of their own learning” (Elmore, 2018, p. 141). The responsibility and control Elmore discussed relate to the concept of agency, defined as an individual’s intention (Bandura, 2006) and capability (Giddens, 1984) to take action in the world to change the course of events (Clarke, Howley, Resnick, & Rose, 2016). It was reported that students high in their sense of agency contributes more to classroom discussion and shows greater learning gains (Clarke et al., 2016). Thus, when designing learning environment, educators need to nurture and promote students’ sense of agency when engaging students in content learning.

In a recent study by Kuhn and colleagues (Kuhn, Iordanou, Matos, Shi, & Hemberger, submitted), it is reported that it is indeed feasible to help students achieve the dual goals outlined
by Elmore (2018). In that study, we engaged students in content learning while engaging them in student-centered and self-directed argument activities. In this study, students gained both argumentation skills and content knowledge following a weeklong dialogue-based argument program centering on a knowledge-dense topic (e.g., whether American workers should enroll in social security system or save on their own).

A dialogue-based argument curriculum

The study reported here utilizes the dialogue-based argument curriculum developed by Kuhn and colleagues in seeking to develop middle school students’ argumentive skills, as well as their intellectual values to engage in deep thinking about complex matters. The curriculum is student-centered, emphasizing peer-to-peer as opposed to teacher-dominated classroom interaction (Kuhn, Hemberger, & Khait, 2016a,b). The curriculum engages students in successive argumentive dialogues with opposing peers on topics of personal and social significance. The dialogue-focused method is grounded in the sociocultural account emphasizing transfer from the inter-mental to the intra-mental level (Vygotsky, 1978; Resnick et al., 2010). Moreover, dialogues provide students with the “missing interlocutor”, without whom students often find the solitary writing activity devoid of purpose (Graff, 2003).

Another prominent feature of the argument curriculum is students’ deep engagement with a topic over multiple class sessions. The rationale is that the development of higher-order intellectual skills, such as the argumentive thinking skills explored here, both warrant and demand sustained and dense practice in which students execute and refine intellectual skills in supportive contexts (Kuhn, Hemberger, & Khait, 2016a).

Significant gains following previous interventions have been identified in individual
argumentive writing (Kuhn & Crowell, 2011; Kuhn, Hemberger, & Khait, 2016b), electronic
dialogic argumentation (Crowell & Kuhn, 2014), face-to-face verbal argumentation (Iordanou,
2013), and meta-level understanding of the norms of argumentation (Kuhn, Zillmer, Crowell, &
Zavala, 2013). Having established the efficacy of the argument curriculum, the next step is to
determine which components of the curriculum contribute to its efficacy and which ones are
expendable, by employing both additive and subtractive research designs.

Zillmer and Kuhn (2018) found that subtracting a consistent same-side peer during
dialogue sessions negatively affected learning outcomes during dialogues. Papathomas and Kuhn
(2017) discovered that adding learning through apprenticeship to dialogue sessions, in which 6th-
grade students engaged in dialogues with an expert adult arguer, accelerated development of
argumentive discourse skills.

Hemberger et al. (2017) also took an additive approach to determine the extent to which
it was possible to accelerate students’ use of evidence in argumentive writing. In previous studies
(Kuhn & Moore, 2014), to provide students with the information they need to engage in deep
discussion of knowledge-dense topics, students were first encouraged to initiate questions on
their own. By the next session, brief factual answers to these questions were made available to all
students. The Q&A method encouraged students’ inquisitiveness about a topic, and helped
students appreciate the potential purpose of the information they obtained. The advantage of the
Q&A over the more traditional method of assigning reading materials on the topic to students
was demonstrated in a recent study (Kuhn et al., submitted). However, Kuhn and Moore (2014)
found that the advantage of participating students over non-participating students did not become
obvious until the second year of the intervention. Even then, the majority of students’ evidence-
based claims demonstrated my-side bias by supporting their own position, and only 3.9% of their
evidence-based claims served the more challenging function of weakening an opposing claim.

To more effectively foster students’ ability to incorporate evidence into their argumentive writing, Hemberger et al. (2017) added to the intervention procedure by providing students with one carefully selected piece of evidence (also in Q&A format) during each of the dialogue sessions, while still encouraging students to ask their own questions. Evidence presented to students followed an optimal sequence based on the differential cognitive demands posed by different types of evidence. Students were first presented with evidence consistent with their position (support my side M+ and weaken other side O-), followed by presentation of the more challenging types of evidence inconsistent with their position (support other side O- and weaken my side M-), as shown in Figure 1. Compared to a comparison condition in which only M+ evidence was presented, and another comparison condition in which no evidence was presented, the intervention condition outperformed both conditions, with 74% of intervention students making weaken-other evidence-based claims at least once in the last intervention topic, significantly higher than the 19% and 0% in the two comparison conditions.

Figure 1. Four argumentive functions of reasons and evidence
Contribution of the present study

Despite the advantages of the modified evidence distribution approach in the Hemberger et al. study (2017), student performance at the end of the one-year intervention was still less than satisfactory. Among essays on the final intervention topic, students included on average only 3 evidence-based claims, roughly one served to support own position (M+) and one weakened the opposing position (O-). Evidence-based claims that supported opposing position (O+) or weakened own position (M-) were scarce.

The low frequency of claims that drew on evidence in general, and of claims that addressed evidence incongruent with students’ own position in particular, should not be simply attributed to students’ lack of access to content knowledge, since a list of topic-related evidence generated during dialogue sessions was made available to students during essay writing. Rather, I would like to propose here, the weakness lies in students’ lack of awareness of the necessity to address evidence, particularly evidence incongruent with their claims, as well as their underdeveloped skills to successfully do so, during argumentative writing.

One central feature of previous interventions following Kuhn’s approach was engaging students in the practice of constructing evidence-based claims. We did not, however, engage students in meta-level reflective exercises that could potentially promote “second order” meta-level understanding regarding the need to address evidence. We hypothesize that the lack of awareness at the meta-level may at least partially account for findings of students’ low frequency of evidence use in argument, particularly claim-incongruent evidence (Kuhn & Moore, 2015; Hemberger et al., 2017). The present study seeks to address this gap, with a study having the goal of enhancing students’ disposition (Kuhn, 2009) and competence to address evidence both congruent and incongruent with their own claims. Targeting development at the meta-level, we
hypothesize that students’ repeated engagement in shared reflection on evidence with a partner will not only highlight for students the importance of constructing evidence-based claims in argumentation, it will also help them develop evaluative criteria to assess the quality of their evidence-based claims and implement them.

The present intervention, emphasizing joint metacognitive reflection, again draws on the sociocultural framework (Vygotsky, 1978; Rogoff, 1990), with the rationale that repeated metacognitive exercise prompted at the external, social level will be interiorized and become operative as well at the individual level (Kuhn, 2000a; Brown, 1997). It is anticipated that when students individually write an argumentive essay at the end of each topic cycle, they will engage in self-reflective and self-regulatory processes, mirroring the reflective and regulatory activities carried out with a partner during their joint sessions. Moreover, when the social support is removed, as is the case for a transfer activity in which students are asked to write individually on a new topic without engaging in dialogue or reflection, reflective exercises conducted with earlier topics will support students’ maintaining meta-level management of their individual behavior, resulting in superior strategic performance in evidence use in written arguments.

In the present study conducted with 7th-grade students from Mainland China, we focus on students’ evidence-based claims in argumentive writing, which served as the principal outcome measure for the present study. The ability to draw on evidence has been identified as a key indicator of argumentive competence (Common Core State Initiative, 2010). Specifically, we seek to test the effectiveness of combining meta-level reflection with practice, over and above practice alone, in fostering students’ disposition and competence to address different kinds of evidence during their argumentive writing. In the remainder of this chapter, we examine research related to individuals’ struggle with arguing from evidence, paying close attention to my-side
bias people exhibit when evaluating evidence. Next, we review research on metacognition and metacognitive training, with particular focus on its role in discourse and its effectiveness in improving student performance. I conclude this chapter by presenting the hypotheses for the present study.

The challenge of arguing from evidence

Distinguish evidence from explanation

Kuhn and Pearsall (2000) identify the essence of mature scientific thinking as the coordination of theory and evidence in a consciously controlled manner. Of particular importance, explanation of the mechanism enabling a cause to achieve its effects should be distinguished from evidence that this is what occurs. Research has consistently demonstrated children’s as well as adults’ struggle with distinguishing causal explanation from evidence.

Kuhn and Pearsall (2000) showed 4-year-olds a sequence of pictures in which two runners competed in a race. Certain pictures provided theory cues as to why one runner won (e.g. he has fancy shoes), while other pictures provided evidence of the outcome (e.g., one runner holds a trophy). To justify their conclusion that one runner won, 4-year-olds were found to struggle with distinguishing between evidence for the claim (holding a trophy) and explanation for it (wearing fancy shoes), as they often referred to the explanation, not evidence, when offering justifications. The struggle is maintained into adulthood as Kuhn (1991) reported that only two-thirds of college-educated participants and one-third of participants without a college education provided genuine evidence to support their theory as to why some students fail in school. Most often, they provided a causal explanation of how this could occur.
Brem and Rips (2000) challenged Kuhn’s (1991) claim that participants did not appreciate the explanation-evidence distinction. They argued that pragmatic factors, such as lack of available evidence, led to reliance on causal explanation. When evidence became available to participants, they demonstrated enhanced competence in referring to evidence during the justification process. Further strengthening their finding, Brem and Rips (2000) reported that participants were metacognitively aware of their strategy of substituting explanations for unavailable evidence.

One limitation in Kuhn’s (1991) and Brem and Rips’s (2000) studies is that both studies were conducted in laboratory settings. Thus it is unknown how people coordinate evidence with claims in more authentic learning situations. Sandoval and Millwood’s study (2005) was different in that they examined high school students’ skill in coordinating evidence with causal claims in a biology classroom in which the concept of evolution was introduced. The authors analyzed written explanations students generated when analyzing scenarios using the theory of natural selection. Specifically, the authors looked at the sufficiency of the evidence students cited for their claims, together with rhetorical quality of the warrants students provided in linking evidence to their claims.

They found high school students to be aware of the need to cite evidence to support claims, in line with Brem and Rips (2000). However, high school students’ rhetorical skill in citing evidence was unsatisfactory, as they often treated evidence as self-evident and failed to interpret the meaning of those data for their arguments. Thus, Sandoval and Milldwood’s study (2005) also supported Kuhn’s (1991) claim that students often failed to distinguish evidence from claim, viewing the two as interchangeable, and thus failed to recognize the need to specify the warrant that explicitly linked evidence to claims.
Overall, these three studies suggest that individuals struggle to distinguish evidence from explanation and the struggle is manifested in two ways. One is individuals view explanation as a substitute for evidence and as a result and do not cite adequate genuine evidence. Another is individuals view evidence and claim as interchangeable and, thus, do not successfully link evidence to claims. Both of these struggles will be observed in our analysis of middle school students’ argumentive writing. Next we examine the confirmation bias in individuals’ evaluation of evidence.

Self-serving evaluation of evidence

The human understanding when it has once adopted an opinion draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, in order that by this great and pernicious predetermination the authority of its former conclusion may remain inviolate. (Bacon, 1620).

The above quote from Francis Bacon from roughly 400 years ago appears accurate, as empirical studies of people’s evaluation of information have amply demonstrated an inherent bias in people’s differential standards in judging opinion-consistent and opinion-inconsistent information (Lord, Ross, & Lepper, 1979; Klaczynski & Gordon, 1996; Edwards & Smith, 1996; Chinn & Brewer, 1993; Kahneman, 2011). Stanovich and West (2007) termed this bias ‘my-side bias,’ since people often interpret and generate evidence in a manner that favors their prior opinions.

One seminal study by Lord, Ross, and Lepper (1979) detected people’s belief polarization following evaluation of mixed or inconclusive evidence. In their study, participants who supported and opposed capital punishment examined two equivalent studies ostensibly offering equivalent levels of support and disconfirmation of the deterrent effects of death
penalty. Rather than moderating their initial views, exposure to inconclusive or mixed evidence widened the gap between their views. The mechanism driving belief polarization, the authors argued, was the biased assimilation process - confirming evidence was readily accepted at face value while disconfirming evidence was subjected to hypercritical scrutiny.

While Lord and colleagues (1979) worked with college students, Klaczynski and Gordon (1996) detected similar belief-maintaining reasoning among adolescents. In their study, Klaczynski and Gordon presented “scientific” evidence relevant to adolescents’ religious affiliation. As in Lord et al. (1979), both belief-consistent and belief-inconsistent evidence were presented. The authors measured participants’ detection of internal validity threat as an indicator of in-depth processing of evidence. The study revealed that belief-inconsistent evidence was subjected to in-depth processing (more threats detected) while belief-consistent evidence to more shallow processing (less threats detected).

Confirmation bias is a pervasive phenomenon (Kuhn, 1989, 1993; Kuhn, Amsel, & O’Loughlin, 1988; Carey, 1986). In fact, a dissociation between general intelligence and reasoning biases has been widely reported (Perkins, 1985; Stanovich & West, 2007, 2008), as participants of both high and low intelligence were found to experience cognitive difficulty in decoupling prior beliefs from the evaluation of belief-relevant evidence and arguments (Stanovich et al., 2013).

While several studies on reasoning have reported participants’ in-depth processing of belief-inconsistent evidence, researchers in argumentation have found that students tended to not address belief-inconsistent evidence when making arguments (Hemberger et al., 2017). These findings are not contradictory however, as the goal of both behaviors is to maintain one’s chosen position.
To maintain one’s chosen position by disregarding belief-inconsistent evidence does not lead to successful argumentative writing. The concept of defeasible reasoning (Pollock, 1987; Hage, 1997; Nussbaum & Edwards, 2011) specifies that the strength of an argument is its ability to withstand refutations. In other words, an argument is warranted if criticisms to it or “alternative viewpoints” are eventually found to be implausible or improbable. Thus, high quality argumentative writing requires writers to introduce plausible objections, in the form of arguments or evidence, that need to be successfully refuted. As O’Keefe (1999) has reported, writings that considered and rebutted counterarguments or counterevidence were more persuasive than texts that did not.

Moreover, several normative models of good or rational thinking involve the ability to decouple from one’s own belief in order to consider and evaluate alternative points of view (Stanovich, West, & Toplack, 2013; Baron, 2008). “Decoupling” is particularly relevant in a digital era where copious unfiltered information is readily accessible to learners of all ages. More than ever, we need learners who are willing and able to critically appraise information both compatible and incompatible with their viewpoints.

Lastly, unbiased consideration of belief-consistent and belief-inconsistent evidence enables authentic conceptual change when teachers seek to help students revise misconceptions. In fact, in the history of science, anomalous data (i.e., data that do not conform to existing theories) have played a crucial role in conceptual revolutions (Thagard, 1992; Dunbar, 1995). Chinn and Brewer (1993) proposed seven distinct forms of response type when individuals are confronted with anomalous data. Some of them constituted authentic theory change while others did not.
To enhance students’ unbiased evaluation of believe-relevant information, as well as their generation of evidence-based arguments that address alternative viewpoints, we propose interventions that target the meta-level. In other words, we need to cultivate learners and writers who are capable of self-monitoring and self-regulating the process of coordinating theory with evidence both consistent and inconsistent with one’s position. In the following section, therefore, we discuss the effectiveness of research focused on metacognitive training.

**Metacognition and metacognitive training**

*Models of metacognition*

Defined in the simplest term, metacognition refers to one’s knowledge and control of one’s own cognition. Research on metacognition stems from Flavell (1979, 1999) who divided metacognition into metacognitive knowledge and metacognitive monitoring and self-regulation. Kuhn (1999, 2000b, 2001), in discussing her model of meta-level knowing, includes three subcomponents: epistemological meta-knowing, declarative meta-level knowing and procedural meta-level knowing. Both models of metacognition incorporate a declarative aspect, which addresses “knowing that”, and a procedural aspect, which addresses “knowing how”, in metacognitive understanding.

*Metacognition in developmental and cognitive research*

Developmental research on metacognition has focused on identifying early origins of meta-level skills. One early origin is 3- to 4-year-olds emerging appreciation of other’s mental states, commonly referred to as Theory of Mind (Perner, 1991; Wellman, 1985, 1990). Since metacognition requires monitoring one’s own and other’s cognitive activity, acquiring a ToM
lays the foundation for more sophisticated metacognitive development later in life (Iordanou, 2016).

Another developmental origin pertains to control at the executive level (Zelazo & Frye, 1998). Zelazo and colleagues, in their simple object-sorting task, reveal three-year-olds’ difficulty in selecting the appropriate strategy to apply under a certain condition, even though both strategies have been acquired and are available. To select and apply an appropriate strategy, an executive function at the meta-level is called upon, one that guides an individual by taking into account the goal of the task as well as understanding of strategies available to oneself (Kuhn & Pearsall, 1998).

As these early appearing metacognitive skills take hold and continue to develop, how do adults perform in tasks involving metacognition? Research on metacognition has traditionally focused on metacomprehension (Thiede, Andersen, & Therriault, 2003) and metamemory (Brown, 1975, 1978; Metcalfe & Kornell, 2005), both of which concern adult participants’ (mostly college-aged) accuracy in their explicit awareness of declarative knowledge gained following a study session that often involve texts or word pairs. While this line of research focuses on the declarative aspect of the models reviewed earlier, another line of research addresses the procedural aspect. Specifically, experiments are conducted to gain insight into how people metacognitively allocate their study time following judgments of learning (Dunlosky & Hertzog, 1998; Metcalfe, 2002; Metcalfe & Kornell, 2005). Metcalfe (2009) reported that participants effectively used metacognition in an effort to selectively study materials in their own region of proximal learning, starting from eliminating materials they have already learned well, then moving on to study easier to more difficult materials.
Metacognitive training in educational research

Lying between early developmental origins and adult-level developmental endpoints, the school years constitute a period of time amenable to intervention. Several studies involving explicit metacognitive training during learning activities have been reported. These studies address different learning goals, ranging from enhancement in argumentative discourse skills (Felton, 2004), argumentive writing (Nussbaum & Edwards, 2011; Song & Ferretti, 2013), mathematical reasoning (Kramarski & Mevarech, 2003), scientific thinking (Zohar & David, 2008), to improvement in problem-solving skills (King, 1991) and reading comprehension (Pressley & Ghatala, 1990).

All these studies point to the efficacy of intervention that combines practice with metacognitive reflection, over and above practice alone. The advantage of metacognitive training is particularly salient during transfer tasks, where gains in strategy performance through practice alone are often lost once the initial instructional context is withdrawn (Kuhn, 2000a).

Nussbaum et al. (2018) tested whether engaging undergraduate students in Walton’s (1996) critical questions would increase students’ integrative refutations in argumentive writing. In both conditions, participants completed a graphic organizer, called an argumentation vee diagram (AVD), designed to elicit arguments for both sides on a social topic. In addition, experimental students were also prompted to answer critical questions that elicit meta-level reflection on reasons for both sides. Experimental and comparison groups did not differ in their generation of arguments and evidence but differed significantly in the number of integrative refutations, defined as statements that effectively countered criticism to one’s own position. Notably, the use of integrative refutation strategies transferred to a writing assignment on a non-intervention topic for the experimental group.
In another study focused on enhancing adolescents’ argumentive discourse strategies, Felton (2004) assigned 7th- and 8th-grade students into either a comparison condition, in which they engaged in argumentative discourse during five weekly sessions, or an experimental condition in which this discourse was coupled with subsequent reflective exercises. The reflective exercise, completed collaboratively by same-side peers, entailed a worksheet designed to elicit reflection on arguments, counterargument and rebuttals in the preceding dialogue. Felton (2004) reported greater gains in the experimental condition in utilizing advanced argumentative discourse strategies. Moreover, while both conditions showed improvement in the intervention topic, only the experimental condition showed transfer of skills to a novel topic.

In a problem-solving study, King (1991) engaged 5th-grade students in computer-assisted programs that make extensive use of graphics and emphasize spatial reasoning (e.g., Rocky’s Boots, Robinett, 1984). King randomly assigned participants to three conditions: guided questioning, unguided questioning, and control. Participants in the guided questioning condition worked in pairs and received a prompt card that listed metacognitive questions they were encouraged to pose to each other in the planning, monitoring, and evaluating phases (e.g., What do we know about the problem so far? Has our goal changed?) Participants in the unguided questioning condition were instructed to ask and respond to each other’s questions but did not receive the prompt card. Those in the control condition received no instruction regarding questioning. Those in the guided questioning condition were not only more successful in solving a novel problem, which spoke to the transferability of acquired skills; they were also more successful in an individual assessment of problem solving, which attested to the internalization of strategies initially obtained in a dyadic, social context.
To account for the superior performance in interventions that combine practice with meta-level reflection, we invoke here the schema theory originating from Schank and Abelson’s (1977) classical paper describing schemas, such as a restaurant schema. A schema is defined as a generic structure that contains conceptual components that need to be instantiated in particular contexts (Nussbaum & Edwards, 2011). Argumentation researchers have adapted the schema theory to describe argument schema, which refers to the mental representation of one’s argumentative knowledge (Reznitskaya et al., 2012).

We propose here that combining practice and meta-level reflection would most effectively help students construct argument schema, which consists of slot-like components that are filled in with particular argumentative moves. Schema reduces cognitive load (Chandler & Sweller, 1991), freeing up more cognitive resources for students to engage in more challenging cognitive activities, such as generating integrative refutations in Nussbaum et al.’s study (2018). Moreover, the “field-invariant” nature of argument schema enables transfer of skill to new argumentative tasks and contexts. Even though a schema is perceived as an internal representation, it is also socially mediated practice (Nussbaum & Edwards, 2011; Greeno, 2006), as all the studies reviewed above engaged students in reflective activities in pair or small-group settings, with gains transferring to individual settings in several studies.

In the present study, by engaging students in joint reflection on evidence both congruent and incongruent with students’ positions, we seek to promote students’ construction of schemas regarding evidence use in argumentation. Moreover, reflective exercises are predicted to help to cultivate self-regulated arguers and writers who are capable of monitoring and regulating their argumentative strategy use, by taking into account argumentation goals that are continuously refined throughout students’ engagement in argument.
Specifically, the proposed hypotheses for the present study are:

H1: For each of the three intervention topics, Evidence Reflection and Argument Practice (ER+AP) students will outperform Argument Practice (AP) students in addressing both claim-congruent and claim-incongruent evidence in argumentative writing.

H2: For transfer essays on a new topic, students in the two Intervention groups will outperform Control students in addressing both claim-congruent and claim-incongruent evidence in argumentative essays, with ER+AP students outperforming AP students.
METHOD

Participants

Participants were 104 students (65 girls and 39 boys) attending school in a large city in Mainland China. Their age ranged from 11 years 10 months to 12 years 10 months. They were entering the seventh grade (beginning of middle school in China) at the start of the intervention. The school was a highly selective private boarding middle school, attended by students from middle to upper middle class Chinese families. Participants were all native-born Chinese, with the Chinese language their first language.

Design

School administrators at the beginning of the school year randomly assigned all entering students to one of eight equivalent advanced classes or one of 14 equivalent regular classes, based on performance on an admissions test administered four months prior to school entry. A cutoff score was used to separate students into advanced and regular levels. Two of the eight advanced classes were randomly chosen by the school officials to participate in the present study. Of the two classes, the researcher randomly chose one class to serve in the Intervention condition (n=54) and the other in the Control condition (n=50). Participants in the Control condition participated only in the pre- and post-assessments and otherwise received their regular instruction. Those in Intervention condition participated in the pre- and post-assessments, and in a four-month argument curriculum designed by the researcher.

The Intervention condition was further divided into an Evidence Reflection and Argument Practice (ER+AP) condition (n=27) and an Argument Practice (AP) condition (n=27). One of the two existing equivalent class subgroups (formed for instruction requiring smaller
groups) was randomly chosen to serve in the ER+AP condition and the other in the AP condition.

To ensure equivalence across conditions prior to the beginning of the intervention, we analyzed results of students’ school diagnostic tests administered during the first week of the school year. The school tests assessed students’ Chinese Language Arts skills and Mathematics skills. Since one of the major outcome measures in the present study is students’ essay writing, the researcher also obtained participants’ score for the essay component of the Chinese Language Arts test. It asked students to write a 600-character narrative essay about a personal experience unique to the student. Table 1 presents means and standard deviations for each test by condition.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>N</th>
<th>Chinese Language Arts</th>
<th>Chinese Language Arts essay component</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER+AP</td>
<td>27</td>
<td>110.30(7.55)</td>
<td>47.85(3.92)</td>
<td>117.26 (17.05)</td>
</tr>
<tr>
<td>AP</td>
<td>27</td>
<td>110.48 (8.40)</td>
<td>46.81(3.39)</td>
<td>117.44 (21.01)</td>
</tr>
<tr>
<td>Control</td>
<td>50</td>
<td>109.96 (6.13)</td>
<td>46.84(3.08)</td>
<td>114.60 (12.40)</td>
</tr>
</tbody>
</table>

Note. Maximum score was 150 for Chinese Language Arts, 150 for Mathematics, and 60 for Chinese Language Arts essay component.

A one-way ANOVA showed no significant difference in essay score across the three conditions, $F (2, 101) = 0.905, p = 0.408$. Nor was there a significant difference in Chinese Language Arts score, $F (2, 101) = 0.051, p = 0.95$, or Mathematics score, $F (2, 101) = 0.375, p = 0.569$.

The intervention constituted a stand-alone course and was referred to as the argument class by school staff and students. Each of the two intervention conditions met three 40-min class
periods per week, for a total of 33 class periods across four months (excluding national holidays and school testing weeks). These 33 class periods normally would have been devoted to Music and Arts classes. The intervention followed that described by Hemberger et al. (2017) with several modifications.

The researcher served as the lead teacher, introduced to students as the “debate coach” in both intervention conditions. This ensued that the instruction given to the two intervention conditions was identical. Two teachers from the school’s staff served as assistant coaches serving in each of the two intervention conditions. To ensure that the two assistant coaches played equivalent roles in the classroom, they participated in a two-hour professional development session provided by the researcher, before the start of the intervention. The PD provided them an overview of the goals of the curriculum and introduction to its activities. The two assistant coaches were not informed of the difference between the two intervention conditions and were thus unaware of the hypotheses of the study. They were instructed that their role in the classroom included handing out and collecting course materials and making sure that students remained engaged in on-task behavior. During the intervention, the researcher observed the two assistant coaches closely and ensured that they played equivalent roles.

**Procedure**

During the week before the start of the intervention, an individual written pre-assessment was administered to students in all conditions during one class period. Students completed these in a whole-class setting, and were given the entire class period (40 minutes) to complete them. Most finished within 30 minutes and all in 40 minutes. Simple clarification of word meaning or task instruction was provided to students if requested, but very few students needed such
clarification. The same assessment was administered again to all students as a post-assessment, during the week following the intervention.

*Pre- and Post-assessment: Individual argumentive essay*

Students were instructed to write a persuasive essay on the topic of whether juveniles who have committed serious crimes should be tried in adult court or juvenile court. The instruction asked students to try to persuade others who hold an opposing view. There was no minimum or maximum word limit. Based on an informal interview with one classroom teacher of these students, it was confirmed that this topic was not addressed in students’ school curriculum and students had not been found to spontaneously engage in discussion of this topic.

A list of 11 pieces of evidence (Appendix A) was distributed to each student before they started writing the essay. Materials in this and subsequent appendices are all English translations of materials presented to students in Chinese. Of the 11 pieces, some tended to support the adult court position and some the juvenile court position. They appeared in a mixed order, constant across students. All students received the same verbal prompt: “Here is some information relevant to the topic, not all information is going to support the side you favor. If it doesn’t, see if you can deal with it. Feel free to use any information you would like in your essay, but you are not required to do so.”

*Intervention topics*

Three topics were chosen for the intervention from an initial set of 10 for which students’ opinions were solicited prior to the start of the intervention. The three chosen were the ones most closely approximating an equal split of students favoring the pro and con side. These appear in Table 2.
Table 2. List of intervention topics

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Should teenagers over 16 only focus on their schoolwork or should they take on a part-time job?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 2</td>
<td>In order to better treat human illnesses, could animals be used to test new medical products and procedures?</td>
</tr>
<tr>
<td>Topic 3</td>
<td>Should the selling of kidney be legalized in China?</td>
</tr>
</tbody>
</table>

*Intervention sequence of activities*

The sequence summarized below occurred in the order shown for each of the three topics. Topic 2 begun at the next class session following completion of the sequence of activities for Topic 1. And Topic 3 followed Topic 2. The average number of weeks during which each topic was engaged was 3 (9 sessions, 3 sessions per week).

Pregame Session 1:

Students were assigned to small groups of 4-5 who shared the same position on the topic. The first class session for each topic involved same-side small group work on generating and evaluating reasons supporting the group’s position. Students were first instructed to write down on blank index cards (3”x5”) reasons for their position. Several blank index cards were provided to each student and they were told to write only one reason per card. Students then worked as a group to share the reasons they generated and rewrite the reasons for better clarity if necessary. They were then asked to decide whether some of the reasons were the same. If so, these reasons should be grouped together. Once cards had been sorted into unique reasons, students were asked to collaborate to evaluate and classify each set of unique reasons into three categories: Best, Good, and So-So.
Game Sessions 2-6:

Students during the game phase were formed into same-side same-gender dyads and the dyad composition stayed intact during the five game sessions. Instant chat software was installed on computers and all discussions between opposing-side dyads were conducted electronically. Each dyad argued with a different opposing-side dyad during each of the five game sessions. Dialogues were monitored from a Master computer to ensure that students stayed on task. If the discussion went off-topic, a reminder to get back on topic was given but no other feedback was provided. Students were instructed to discuss with their partner before writing and only do so when they had agreed on what to say. While awaiting response from the opposing pair, dyads were asked to complete a sheet designed to promote reflection on the dialogues. These were of two alternating forms. Students began with identifying and reflecting on the opponents’ main argument, and then reflecting on their own side’s main argument (Appendix B), with the two forms alternating across sessions.

During game sessions, students were also encouraged to ask questions regarding the topic for information they would like answered to help them make their arguments. Students were asked to write down their question and place it in a designated space. Answers to these questions were provided and circulated among all students from both intervention conditions at the following session.

In addition to answers to students’ self-generated questions, each dyad also received a short piece of information, in question-and-answer format, potentially usable during each of the five game sessions. The sequence of five pieces of potential evidence (one per game session) followed this order:

1. Potential evidence functioning to support student’s favored position on the topic (M+)
2. Potential evidence functioning to weaken opposing position on the topic (O-)

29
3. Potential evidence functioning to support opposing position on the topic (O+)
4. Potential evidence functioning to weaken student’s favored position on the topic (M-)
5. Second instance of type 1 (M+)

Students on opposing sides thus received a different sequence of information, but students on the same side received an identical sequence. The support-own (M+) type was presented first as previous research had found it to be easiest for students. The next piece of evidence given was the weaken-other (O-) type, which posed a greater challenge given novice students tend to neglect the opposing position. The most challenging types of evidence are those incongruent with individual’s own position (support-other O+ and weaken-own M-). They were presented during the next two sessions. In the fifth session, students encountered a second piece of support-own evidence. Sets of evidence presented for each topic are shown in Appendix C.

Showdown Session 7:

During the showdown session, students reassembled in their same-side small groups and took turns in occupying the “hot-seat” at the center of the room, where a student from each side verbally debated student from the opposing side. Each round lasted for 3 minutes, after which a new pair of opposing-side students took center stage. A huddle, lasting 1 minute, could be called either by the speaker in the hot-seat or members of the speaker’s team. During a huddle, both speakers rejoined their respective teams and received suggestions from their peers before the debate resumed. On average, there were 7 rounds of debate during a showdown session.

Showdown Debrief Session 8:

All showdown sessions were video recorded and subsequently transcribed by the researcher to produce a showdown map. A strategy code was assigned to each turn in the
showdown map based on its argumentive quality, most specifically its functional relation to the immediately preceding opposing claim, and its success in advancing the argumentive objectives of weakening an opposing claim and advancing one’s own claims. Successful moves (such as counterargument or evidence use) were assigned a positive point and unsuccessful moves (such as unwarranted assumption or unconnected statement) received a negative point. The lead teacher reminded students of these strategic goals at the beginning of each debrief session.

During the debrief session, each student received a copy of the showdown map and reviewed the scoring with the lead teacher. Students were allowed to question the scoring on the transcript and encouraged to elaborate the reasons for their disagreement, and whole-class discussion followed. At the end of the debrief session, positive and negative points were summed, total scores calculated, and a winning side declared.

Individual Essay Session 9:

The culminating activity for each topic cycle was students’ composition of a “letter to a newspaper editor” justifying why their position on the topic was better than the opposing one. Students were told that the goal of their writing was to try to persuade readers who held an opposing opinion. Before writing, students received a list of evidence that they have encountered earlier to help with their writing (Appendix D). The same verbal prompt given during pre-assessment was given again: “Here is some information relevant to the topic, not all information is going to support the side you favor. If it doesn’t, see if you can deal with it. Feel free to use any information in your essay you would like, but you are not required to do so.” Session 9 was the only session in which students worked individually. The next session marked the beginning of the activity cycle for the next topic.
Condition manipulation

The procedure for the ER+AP condition and AP condition was identical except for one respect. In the ER+AP condition, in addition to the argument reflection sheet described earlier, dyads were asked to collaboratively complete two parts of a sheet designed to elicit shared reflection on evidence. One addressed anticipation of potential evidence use before dialogue with the opposing side, and the other involved evaluation of actual evidence use toward the end of the dialogue. The AP condition, in contrast, did not include these evidence-related reflective exercises. They were only asked to try to use the evidence in dialogues.

In the ER+AP condition, before the start of each dialogue session, the dyad was presented the evidence for that session, and along with it, Part A of the Evidence Reflection Sheet (shown below). Dyads were asked to discuss with their partner before writing an answer for each question. Electronic dialogues began 4 minutes later, and dyads continued to complete Part A while they awaited response from the opposing pair. Twenty-five minutes into the chat session, Part B of Evidence Reflection Sheet and argument reflection sheet were distributed to all dyads. Dyads were asked to decide the order in which these were completed. Again, dyads were instructed to discuss and agree before writing down their answers.

<table>
<thead>
<tr>
<th>Evidence Reflection Sheet – Part A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructions: the following questions will help you think about the evidence you have today. Please discuss with your partner before answering each question.</td>
</tr>
</tbody>
</table>

1. Have you heard this evidence from your opponents before? (Please circle one)
   - Yes
   - No
   - Not sure

2. Does the evidence help your side or your opponent’s side? (Please circle one)
   - Help my side
   - Help opponent’s side
   - Not sure
If it helps your side, answer questions 3 & 4. If it helps your opponent’s side, answer questions 4 & 5. If you are not sure, answer question 4.

3. Since this evidence helps your side, is there anything your opponents might say against it?

4. How might someone who disagree with you on this issue use this evidence?

5. Since this evidence helps your opponent’s side, is there anything you can say against it?

Evidence Reflection Sheet - Part B

1. Have you addressed this evidence today? If yes, what argument did you make with it? If no, why?

2. Are you satisfied with the way you and your partner used the evidence today?
   
   Satisfied  So-so  Not satisfied

3. If you are satisfied, explain why you are satisfied.

   If you are not satisfied, why? Can you think of a way to improve how you used it?
RESULTS

The results section consists of two parts. The first part involves quantitative and qualitative analysis of intervention topic essays and transfer topic essays. The second part traces developmental patterns in students’ responses to the evidence reflection prompts.

For intervention topic essays, it should be noted that the final essays were students’ only individual work for that topic (a fact that raises concern with independence of observations in statistical analysis), preceded by small-group and dyadic collaborative activity that we perceive as essential in developing students’ argument skills. A student’s final essay thus constituted a representation of what he or she had taken away from this extended collaborative engagement with the topic.

To assess the extent to which writings of students who worked in a dyad resemble each other, we used the variance components model to assess intraclass correlation (ICC) for each of the outcome measures. Since dyad composition changed from topic to topic, we examined intraclass correlation within each of the three topics. For Topic 1, we obtained the following values for each of the measures: Idea units (ICC = 0), Functional evidence-based units (ICC = 0.020), Claim-congruent functional evidence-based units (ICC = 0.043), Claim-incongruent functional evidence-based units (ICC = 0.158), Evidence-based However units (ICC = 0), Full evidence-based However units (ICC = 0). For Topic 2, the ICC values for the above measures ranged between 0 and 0.245. For Topic 3, the ICC values for the same measures ranged between 0 and 0.254. Thus, we conclude that intraclass correlation within dyads wasn’t a significant concern for the present study. For the transfer topic, since students did not engage in any form of collaborative activity before writing individual essays, intraclass correlation was not an issue.
Coding of essays

Since all the essays were written in the Chinese language, the coding took place in Chinese to avoid distortion or loss of meaning due to translation. A Master’s student who recently graduated from Teachers College and who speaks Chinese as a native language was recruited to establish coding reliability with the researcher. The coding assistant had background knowledge of the argument curriculum but was unaware of the experimental setup and hypotheses of the study. I first introduced the essay coding scheme to the assistant and answered all her questions. The two coders then independently coded three essays, compared results, and had an in-depth discussion and resolved differences in coding. These practice trials were repeated four times before the two coders achieved at least 85% agreement in essay coding. Blind to condition, the two coders then moved on to independently code 20% of the essay dataset, totaling 76 essays randomly selected from the dataset.

Each essay was first segmented into idea units, defined as a claim with supporting argument and/or evidence that intended to convey a specific single point. In the present study, the number of idea units in an essay is treated as an indicator of essay length. The two coders segmented each essay, calculated the total number of idea units for each essay, and reached an agreement of 88%. The two coders then had a discussion and resolved all differences in segmentation of essays. Each idea unit was further coded as either evidence-based or non-evidence-based. An evidence-based unit was defined as one that drew on evidence, either from a shared source provided by the researcher or from students’ personal knowledge source. A non-evidence-based unit was defined as one that did not make any reference to evidence.

Evidence-based units were further classified into two categories, functional use of evidence and nonfunctional use of evidence, based on the coding scheme developed and utilized
in one of our previous studies (Hemberger et al., 2017). *Functional* use of evidence referred to successful use of evidence in service of a claim, in which students explicitly connected evidence and the claim it intended to support. *Nonfunctional* use of evidence referred to unsuccessful use of evidence in service of a claim, which included mischaracterized evidence, missing or unsuccessful connection of evidence to a claim. Figure 2 provides a diagrammatic representation of the essay coding scheme.

Figure 2. Diagrammatic representation of the essay coding scheme

![Diagram of essay coding scheme]

Table 3 presents a detailed illustration of each of the categories as well as translated verbatim examples of students’ evidence-based claims coded into each category. Based on the argumentive goal a claim served, *functional* evidence-based claims were further divided into four categories, support-own side (M+), weaken opposing side (O-), support opposing side (O+) and weaken own side (M-). In later analyses, we collapsed M+ and O- units as claim-congruent units since the argumentive goal is consistent with students’ own claims. We collapsed O+ and M-
units as claim-incongruent units since the argumentive goal is inconsistent with students’ own claims. Table 4 presents translated verbatim examples of functional evidence-based units coded into each category.

To summarize, having segmented an essay into idea units, two coders assigned each idea unit into one of six categories: (1) non-evidence-based unit, (2) non-functional evidence-based unit, (3) functional evidence-based unit that supports own side, (4) functional evidence-based unit that weakens opposing side, (5) functional evidence-based unit that supports opposing side, (6) functional evidence-based unit that weakens own side. Two coders achieved an inter-rater agreement of 84% in assigning an idea unit into one of the six categories, Cohen’s kappa = .73, p < .005. Discrepancy was resolved through discussion and the researcher proceeded to code the remaining essays.

Table 3. Coding scheme for categorizing evidence-based units

<table>
<thead>
<tr>
<th>Level</th>
<th>Category definition</th>
<th>Verbatim (translated) examples*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional evidence-based units</td>
<td>Explicitly and successfully used evidence in service of a claim</td>
<td>Selling of kidney should be legalized because we do not have enough kidneys at the moment. In fact, in China in 2012, for every 150 people waiting for a donated kidney, 149 die while waiting.</td>
</tr>
<tr>
<td>Nonfunctional evidence-based units</td>
<td>Evidence mischaracterized in a way that substantively misrepresents its meaning</td>
<td>We don’t need to legalize the kidney sale market because we don’t have that many patients who need a kidney. In 2012, we only have 150 people who are waiting for a donated kidney.</td>
</tr>
<tr>
<td></td>
<td>Missing connection of evidence to claim. Usually partial or complete copy of evidence</td>
<td>Statistics show that in China in 2012, for every 150 people waiting for a donated kidney, 149 die while waiting.</td>
</tr>
<tr>
<td></td>
<td>Attempted but unsuccessful connection of evidence to claim</td>
<td>The information says that 149 out of every 150 people die waiting for a donated kidney. The information suggests that kidney is just a commodity.</td>
</tr>
</tbody>
</table>

*Note. All the examples refer to the following piece of evidence in Topic 3

Q: Do people die because they can't get a new kidney in time?
A: Yes, statistics show that in China in 2012, for every 150 people waiting for a donated kidney, 149 die while waiting.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Support own side (M+)</td>
<td>Data show that drugs that come out of animal testing can be used on sick animals as well. So this is a win-win situation in which both humans and animals benefit from animal testing.</td>
<td>If we don’t use animals for testing, we could simply use dead bodies to determine the cause of death and to help us improve our drugs.</td>
</tr>
</tbody>
</table>
|                                | [Q: Can medical testing of animals be of any benefit to animals?  
A: Many of the medications that are given to sick animals (such as pets and zoo animals) were discovered as a result of medical research with humans that involved those animals.]* | [Q: Can bodies of humans who recently died be used for research?  
A: Examining human bodies soon after death can help to better understand causes and effects of diseases and medicines.] |
| Weaken opposing side (O-)      | If we don’t allow animal testing, then a lot of advanced studies such as those involving gene engineering can’t be carried out. | According to the statistics of the Food and Drug Administration, 92 out of every 100 drugs that pass animals tests fail in humans. The low success rate means we waste too animals on producing drugs that do not actually improve our drugs. |
|                                | [Q: Are there any types of research that could be performed with animals but not humans?  
A: Many studies of living bodies are so complicated and uncertain that they could only be carried out with animals. For example, studies in gene engineering test how to modify the organs of animals so they can be transplanted to humans.] | [Q: Do most of the drugs that pass animal tests succeed in humans?  
A: The Food and Drug Administration reports that 92 out of every 100 drugs that pass animal tests fail in humans.] |
| Support opposing side (O+)     | It is true that there are alternatives to using animals in research. For instance, synthetic human skin be can be used to test the effects of sunscreen. | Admittedly, there are laws in place to limit what scientists could do with animals in their laboratories. |
|                                | [Q: Can synthetic versions of human organs be used in research?  
A: Studies involving the effect of sunscreen on a material like human | [Q: How are animals treated in research laboratories?  
A: There are laws in place to help ensure that distress and pain in animals is kept to a minimum, but |
skin gave quick results, compared to the length of time required for animal testing.

the daily treatment of animals is not known because the testing places cannot be monitored at all times and records are not shared.

Weaken own side (M-)
The opposing side made a valid point that not all drugs that pass animal testing would be successful on humans

[Q: Do most of the drugs that pass animal tests succeed in humans?
A: The Food and Drug Administration reports that 92 out of every 100 drugs that pass animal tests fail in humans.]

It is true that if scientists did not experiment on animals, we would not have the drugs available to treat diabetes, hepatitis, polios, and AIDS.

[Q: Has animal testing led to cures for any human diseases?
A: Animal testing has led to treatments and cures for many human diseases. For example, research with dogs led to treatments for diabetes, and research with monkeys have led to treatments for hepatitis, polio, and AIDS.]

*Note. Evidence students used are shown in brackets.

Given our interest in understanding students’ ability to construct integrative statements that connect two conflicting idea units, we took a further step by re-examining all the essays and coded students’ generation of However units. A However unit is a superordinate category that integrates two adjacent idea units that serve contrasting argumentive functions, with the first unit either supporting the opposing side or weakening own side, followed by the second unit either weakening the opposing side or supporting own side, as diagrammatically illustrated in Figure 3. Thus, a However unit can be conceived as a strategic sequence intended to restore the strength of one’s position following introduction of arguments and/or evidence unfavorable to one’s position.
For instance, the following statement from a student against animal research qualified as a *However* unit, “Even though a lot of people consider rats as useless creatures, I think every creature that continues to exist over the past thousands or millions of years does so for a good reason.” Note that the first statement conceded to an alternative viewpoint that rats are “useless” (thus they can be used for research), and the second statement introduced a conflicting statement that their continued existence should be for a good reason (thus they should not be used for research).

Given our main focus on students’ ability to draw on evidence in writing, we assigned each *However* unit into one of two categories: evidence-based or non-evidence-based. An evidence-based *However* unit makes reference to evidence whereas a non-evidence-based *However* unit does not. Among evidence-based *However* units, we singled out those that successfully connected two pieces of conflicting evidence, and labeled them as full evidence-based *However* units. For example, the following statement from a student against animal research qualified as a full evidence-based *However* unit: “Some people argue that animal testing also benefits animals as they could receive better treatments when they get sick. However, statistics show that 1.2 million animals were used in research in 2005, so the amount of animals actually benefitting from research may not be proportional to the number of animals used in...
research.” Note that both the first part and second part of this *However* unit made reference to evidence.

Two coders independently coded the same 76 essays for *However* units and achieved an inter-rater agreement of 97%, Cohen’s kappa = .883, $p < .0005$. Differences were resolved through discussion and the researcher proceeded to code the remaining essays.

*Analysis of intervention topic essays*

A Generalized Linear Model (GLM) with negative binomial regression was carried out within each of the three intervention topics to compare across the Evidence Reflection and Argument Practice (ER+AP) condition and Argument Practice (AP) condition in the mean number of functional evidence-based units, claim-congruent functional evidence-based units and claim-incongruent functional evidence-based units. We chose negative binomial regression as a tool for statistical analysis since it is appropriate for modeling count outcome variables. Before selecting negative binomial regression, we also fit the Poisson regression model. Goodness-of-fit statistics suggest that negative binomial regression was a better fit than Poisson regression for our dataset.

During Topic 1, as shown in Figure 4, the two conditions did not significantly differ in the mean number of functional evidence-based units, with 4.04 ($SD = 1.91$) for the ER+AP condition and 3.69 ($SD = 2.60$) for the AP condition, $p = .526$. We broke down functional evidence-based units into claim-congruent and claim-incongruent types, which will be analyzed separately.

The mean number of claim-congruent functional evidence-based units was 3.27 ($SD = 1.89$) for the ER+AP condition and 3.50 ($SD = 2.61$) for the AP condition, a non-significant difference, $p = .651$. As shown in Table 5, the percentage of students who ever made a claim-
congruent functional evidence-based claim was 89% for the ER+AP condition and 92% for the AP condition, a non-significant difference (Fisher’s Exact Test, \( p = 1.00 \)).

The two conditions did differ significantly, however, in the mean number of claim-incongruent functional evidence-based units, with a mean of \( .77 (SD = .76) \) for the ER+AP condition and a mean number of \( .19 (SD = .40) \) for the AP condition. The ER+AP condition had an expected 4.00 (95% CI, 1.501 to 10.658) times more claim-incongruent functional evidence-based units than the AP condition, a significant result, Wald \( \chi^2 (1) = 7.687, p = .006 \). The percentage of students who ever made a claim-incongruent functional evidence-based claim was 62% for the ER+AP condition and 19% for the AP condition, a significant difference (Fisher’s Exact Test, \( p = .004 \)). To conclude, in Topic 1, both group mean analysis and individual-level analysis indicated that the ER+AP condition and the AP condition did not differ much in using claim-congruent evidence, but the RP condition addressed significantly more evidence incongruent with own position.
Figure 4. Mean number of different types of evidence-based units by condition at Topic 1.

Table 5. Percentage of students who ever made a certain type evidence-based claim by condition and topic

<table>
<thead>
<tr>
<th></th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ER+AP condition (n=26)</td>
<td>AP condition (n=26)</td>
<td>ER+AP condition (n=27)</td>
</tr>
<tr>
<td>Claim-congruent functional</td>
<td>89%</td>
<td>92%</td>
<td>100%</td>
</tr>
<tr>
<td>Claim-incongruent functional</td>
<td>62%</td>
<td>19%</td>
<td>78%</td>
</tr>
<tr>
<td>Functional</td>
<td>96%</td>
<td>92%</td>
<td>100%</td>
</tr>
</tbody>
</table>

By Topic 2, as shown in Figure 5, the two conditions did not significantly differ in the mean number of any of the following variables: functional evidence-based units, claim-congruent functional evidence-based units, and claim-incongruent evidence-based units. As
presented in Table 5, the percentage of students who ever made a functional evidence-based claim, claim-congruent functional evidence-based claim and claim-incongruent functional evidence-based claim also did not differ significantly across conditions. Thus, group mean analysis and individual-level analysis indicated that the two conditions showed equivalent performance in Topic 2 essays.

Figure 5. Mean number of different types of evidence-based units by condition at Topic 2

By Topic 3, as shown in Figure 6, a significant difference was found in the mean number of functional evidence-based units for the ER+AP condition \((M = 6.19, SD = 2.40)\) and the AP condition \((M = 3.37, SD = 1.57)\). The ER+AP condition had an expected 1.837 (95% CI, 1.421 to 2.376) times more functional evidence-based units than the AP condition, a significant result, Wald \(\chi^2 (1) = 21.512, p < .0005\).

For claim-congruent functional evidence-based units, the mean number was 4.62 \((SD = 1.94)\) for the ER+AP condition and 3.04 \((SD = 1.45)\) for the AP condition. The ER+AP condition
had an expected 1.520 (95% CI, 1.148 to 2.012) times more claim-congruent evidence-based units than the AP condition, a significant result, Wald $\chi^2 (1) = 8.532, p = .003$. The proportion of students who ever made a claim-congruent functional evidence-based statement was 96% for the ER+AP condition and 96% for the ER+AP condition, a non-significant difference (Fisher’s Exact Test, $p = 1.00$).

For claim-incongruent functional evidence-based units, the mean number was 1.65 ($SD = 1.29$) units for the ER+AP condition and .37 ($SD = .56$) units for the AP condition. The ER+AP condition had an expected 4.465 (95% CI, 2.244 to 8.886) times more claim-incongruent evidence-based units than the AP condition, a significant result, Wald $\chi^2 (1) = 18.166, p < .0005$. The proportion of students who ever made a claim-incongruent functional evidence-based statement was 81% for the ER+AP condition and 33% for the AP condition, a significant difference (Fisher’s Exact Test, $p = 0.001$). Thus, in Topic 3, group mean analysis and individual-level analysis indicated that the ER+AP condition outperformed the AP condition in successfully addressing evidence both congruent and incongruent with own position.
The above analysis did not control for essay length, which was found to vary from topic to topic. At Topic 1, the mean number of idea units was 15.85 ($SD = 4.21$) for the ER+AP condition and 17.19 ($SD = 4.27$) for the AP condition. At Topic 2, the mean number of idea units was 14.67 ($SD = 3.61$) for the ER+AP condition and 15.81 ($SD = 5.86$) for the AP condition. At Topic 3, the mean number of idea units was 12.42 ($SD = 3.51$) for the ER+AP condition and 13.44 ($SD = 4.49$) for the AP condition. Negative binomial regression indicated that none of the condition differences in the mean number of idea units reached significance at each topic.

To control for essay length, we calculated each essay’s percentage of claim-congruent functional evidence-based units, and percentage of claim-incongruent functional evidence-based units, out of total idea units. To get a fuller picture of how groups progressed from Topic 1 to
Topic 3, we conducted a two-way mixed ANOVA with percentage as an outcome variable, condition as a between-subjects factor and topic as a within-subjects factor.

As shown in Figure 7, for claim-congruent functional evidence-based claims, there was a statistically significant two-way interaction between condition and topic, $F(2, 98) = 3.213, p = 0.045$, Partial $\eta^2 = .101$. Analysis of simple main effect for condition was carried out to enable comparison at each topic. At Topic 1, the mean percentage of claim-congruent functional evidence-based claims was 21% ($SD = 0.12$) for the ER+AP condition and 20% ($SD = 0.14$) for the AP condition, a non-significant difference, $F(1, 50) = 0.019, p = 0.890$, Partial $\eta^2 < 0.0005$. At Topic 2, the mean percentage was 36% ($SD = 0.11$) for the ER+AP condition and 28% ($SD = 0.14$) for the AP condition, a significant difference, $F(1, 52) = 4.846, p = 0.032$, Partial $\eta^2 = .085$. At Topic 3, the mean percentage was 38% ($SD = 0.14$) for the ER+AP condition and 25% ($SD = 0.14$) for the AP condition, a significant difference, $F(1, 51) = 11.971, p = 0.001$, Partial $\eta^2 = .190$. Thus, the ER+AP condition had a significantly higher percentage of claim-congruent functional evidence-based claims at both Topic 2 and Topic 3.

Examination of simple main effects for topic was also carried out to see whether each condition showed significant change across time. For the ER+AP condition, there was a statistically significant effect of time on the mean percentage of claim-congruent functional evidence-based claims, $F(2, 48) = 13.567, p < 0.0005$, Partial $\eta^2 = .402$. For the AP condition, the effect of time on the mean percentage of claim-congruent functional evidence-based claims approached significance, $F(2, 50) = 3.134, p = 0.052$, Partial $\eta^2 = .111$.

For claim-incongruent functional evidence-based claims, there was also a statistically significant two-way interaction between condition and topic, $F(2, 98) = 7.595, p = 0.001$, Partial $\eta^2 = .170$. Analysis of simple main effect for condition was carried out to enable comparison at
each topic. At Topic 1, the mean percentage of claim-incongruent functional evidence-based claims was 5% ($SD = 0.05$) for the ER+AP condition and 1% ($SD = 0.03$) condition, a significant difference, $F (1, 50) = 11.630, p = 0.001$, Partial $\eta^2 = .189$. At topic 2, the mean percentage was 9% ($SD = 0.07$) for the ER+AP condition and 6% ($SD = 0.06$) for the AP condition, a significant difference, $F (1, 52) = 4.369, p = 0.042$, Partial $\eta^2 = .078$. At topic 3, the mean percentage was 13% ($SD = 0.09$) for the ER+AP condition and 2% ($SD = 0.04$) condition, a significant difference, $F (1, 51) = 30.963, p < 0.0005$, Partial $\eta^2 = .378$. Thus, even though both conditions had lower percentages of claim-incongruent evidence-based claims compared to those of claim-congruent evidence-based claims, the ER+AP condition still showed significant higher percentages for all three topics.

Examination of simple main effect for topic was also carried out to see whether each condition showed significant change across time. For the ER+AP condition, there was a significant effect of time on the mean percentage of claim-incongruent functional evidence-based claims, $F (2, 48) = 9.780, p < 0.0005$, Partial $\eta^2 = .290$. For the AP condition, there was also a significant effect of time on the mean percentage of claim-incongruent functional evidence-based claims, $F (2, 50) = 8.479, p = 0.001$, Partial $\eta^2 = .253$. In conclusion, analysis based on mean percentage yielded a slightly different picture than previous analysis based on mean frequency. While analysis on mean frequency did not showed condition difference at Topic 2, analysis based on mean percentage detected significant condition difference for both claim-congruent and claim-incongruent functional evidence-based claims.
Figure 7. Mean percentage of idea units served as claim-congruent and claim-incongruent functional evidence-based claims by condition and topic

In addition, we examined frequencies of evidence-based *However* statements in essays, as presented in Figure 8. At Topic 1, the mean number of evidence-based *However* statements was 1.69 ($SD = 1.52$) for the ER+AP condition and 0.65 ($SD =1.13$) for the AP condition. The ER+AP condition had an expected 2.588 (95% CI, 1.323 to 5.063) times more evidence-based *However* units than the AP condition, a significant result, $Wald \chi^2 (1) = 11.089, p = 0.005$. As shown in Table 6, at Topic 1, the proportion of students who ever made an evidence-based *However* unit was 73% for the ER+AP condition and 39% for the AP condition, a significant difference (Fisher’s Exact Test, $p = 0.012$).
Figure 8. Mean number of evidence-based *However* units and full evidence-based *However* units by condition and topic

![Graph showing mean number of evidence-based *However* units and full evidence-based *However* units by condition and topic.]

*Note.* **p<.01, ***p<.001

Table 6. Percentage of students who ever showed evidence-based *However* units and full evidence-based *However* units by condition and topic

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ER+AP condition</strong></td>
<td><strong>AP condition</strong></td>
<td><strong>ER+AP condition</strong></td>
</tr>
<tr>
<td>Evidence-based <em>However</em></td>
<td>73%</td>
<td>39%</td>
</tr>
<tr>
<td>Full evidence-based <em>However</em></td>
<td>8%</td>
<td>4%</td>
</tr>
</tbody>
</table>

We further elevated the criteria and examined the mean number of evidence-based *However* units with both parts drawing on evidence, termed full evidence-based *However* units.
As presented in Figure 8, the mean number was .08 (SD = .27) for the ER+AP condition and 0.04 (SD = .02) for the AP condition, a non-significant difference, \( p = .571 \). The proportion of students who ever made a full evidence-based *However* was 8% for the ER+AP condition and 4% for the AP condition, a non-significant difference (Fisher’s Exact Test, \( p = 1.000 \)). To conclude, at Topic 1, ER+AP students constructed significantly more evidence-based *However* units. ER+AP students were not any more successful in constructing full evidence-based *However* units, presumably because the skills involved were more complex and took longer to develop.

At Topic 2, the two conditions did not show significant difference in the mean number of evidence-based *However* units and full evidence-based *However* units. Percentages of students who ever included the above category also did not significantly differ across conditions.

At Topic 3, the mean number of evidence-based *However* units was 2.81 (SD = 2.19) for the ER+AP condition and 0.67 (SD = 0.88) for the AP condition. The ER+AP condition had an expected 4.212 (95% CI, 2.514 to 7.054) times more evidence-based *However* units than the AP condition, a significant result, Wald \( \chi^2 \) (1) = 29.852, \( p < .0005 \). The proportion of students who ever made an evidence-based *However* unit was 85% for the ER+AP condition and 52% for the AP condition, a significant difference (Fisher’s Exact Test, \( p = 0.011 \)).

There was also a significant difference in full evidence-based *However* units between the ER+AP condition (\( M = 1.65, SD = 1.57 \)) and the AP condition (\( M = 0.26, SD = 0.53 \)) at Topic 3. The ER+AP condition had an expected 6.379 (95% CI, 2.862 to 14.219) times more full evidence-based *However* units than the AP condition, a significant result, Wald \( \chi^2 \) (1) = 20.671, \( p < .0005 \). The proportion of students who ever made a full evidence-based *However* unit was 65% for the ER+AP condition and 22% for the AP condition, a significant difference (Fisher’s
Exact Test, \( p = .002 \). In conclusion, in the last intervention topic, the ER+AP condition had demonstrated significant gains in the challenging task of connecting two contradictory pieces of evidence, to form what we termed full evidence-based *However* units.

*Analysis of transfer essays*

The number of essays analyzed at pretest was 26 for the ER+AP condition, 27 for the AP condition and 50 for the Control condition. The number of essays analyzed at posttest was 27 for the ER+AP condition, 27 for the AP condition and 50 for the Control condition. Reduced numbers were due to student absences for medical reasons. We first ran a negative binomial regression for pretest essays to establish equivalence across the three conditions before the intervention started. Negative binomial regression with condition as the predictor variable showed that the three conditions did not differ significantly in the following measures at pretest: idea units, functional evidence-based units, claim-congruent functional evidence-based units, claim-incongruent functional evidence-based units, evidence-based *However* units, and full evidence-based *However* units.

Having established equivalence across conditions at pretest, we now use negative binomial regression with condition as a predictor variable to compare performance across conditions at posttest. We also controlled for pretest performance by treating it as a covariate. After adjusting for pretest performance, there was no significant difference in the mean number of idea units at posttest (Table 7), \( \text{Wald} \chi^2 (2) = 0.579, p = 0.749 \). Thus, posttest difference across conditions in other measures could not simply be attributed to differences in essay length.

At posttest, we found a significant difference in the mean number of functional evidence-based idea units across the three conditions (Table 7), \( \text{Wald} \chi^2 (2) = 67.314, p < 0.0005 \). The
ER+AP condition had an expected 1.706 (95% CI, 1.344 to 2.166) times more functional evidence-based units than the AP condition, a significant result, $\chi^2(1) = 19.270, p < 0.0005$. The AP condition had an expected 1.470 (95% CI, 1.142 to 1.892) times more functional evidence-based units than the Control condition, a significant result, $\chi^2(1) = 8.969, p = 0.003$.

Table 7. Descriptive statistics for measures in pre- and posttests by condition

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>ER+AP condition</th>
<th>AP condition</th>
<th>Control condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD) idea Units</td>
<td>Pretest</td>
<td>8.58(2.12)</td>
<td>9.30(3.11)</td>
<td>9.34(3.37)</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>12.04(3.04)</td>
<td>11.93(4.20)</td>
<td>11.66(2.77)</td>
</tr>
<tr>
<td>Mean (SD) functional evidence-based units</td>
<td>Pretest</td>
<td>3.46(1.48)</td>
<td>3.22(2.01)</td>
<td>3.34(1.77)</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>6.85(2.89)</td>
<td>4.00(1.59)</td>
<td>2.74(1.64)</td>
</tr>
</tbody>
</table>

For claim-congruent functional evidence-based units (Figure 9), at posttest, the mean number was 5.78 ($SD = 2.59$) units for the ER+AP condition, 3.52 ($SD = 1.55$) units for the AP condition, and 2.38 ($SD = 1.41$) for the Control condition, a significant difference across the three conditions, $\chi^2(2) = 55.046, p < 0.0005$. The ER+AP condition had an expected 1.635 (95% CI, 1.265 to 2.113) times more claim congruent functional evidence-based units than the AP condition, a significant result, $\chi^2(1) = 14.125, p < 0.0005$. The AP condition had an expected 1.509 (95% CI, 1.152 to 1.977) times more claim congruent functional evidence-based units than the Control condition, a significant result, $\chi^2(1) = 8.918, p = 0.003$. 
For claim-incongruent functional evidence-based units (Figure 10), at posttest, the mean number was 1.04 ($SD = 0.85$) units for the ER+AP condition, 0.48 ($SD = 0.64$) units for the AP condition, and 0.36 ($SD = 0.60$) for the Control condition, a significant difference across the three conditions, Wald $\chi^2 (2) = 13.285, p = 0.001$. The ER+AP condition had an expected 2.183 (95% CI, 1.125 to 4.234) times more claim-incongruent functional evidence-based units than the AP condition, a significant result, Wald $\chi^2 (1) = 5.329, p = 0.021$. The AP condition and Control condition did not differ significantly in the mean number of claim-incongruent functional evidence-based units, Wald $\chi^2 (1) = 0.559, p = 0.455$. In conclusion, at posttest, the ER+AP condition outperformed the other two conditions in making both claim-congruent and claim-incongruent evidence-based claims. The AP condition outperformed the Control condition only in the easier task of making claim-congruent evidence-based claims.
Since all three conditions wrote longer essays at posttest (Table 7), to control for essay length at pretest and posttest, we also calculated each essay’s percentage of claim-congruent functional evidence-based units, as well as percentage of claim-incongruent functional evidence-based units, out of total idea units. For the ER+AP condition, the percentage of claim-congruent functional evidence-based units increased from 37% (SD = 0.15) at pretest to 48% (SD = 0.18) at posttest, a significant increase as indicated by a one-way repeated measures ANOVA, $F(1, 25) = 11.585, p = 0.002$, Partial $\eta^2 = .317$. The percentage of claim-incongruent functional evidence-based units increased from 4% (SD = 0.08) at pretest to 8% (SD = 0.06) at posttest, a significant increase, $F(1, 25) = 5.573, p = 0.026$, Partial $\eta^2 = .182$.

For the AP condition, the percentage of claim-congruent functional evidence-based units was 29% (SD = 0.18) at pretest and 31% (SD = 0.15) at posttest, a non-significant change. The percentage of claim-incongruent functional evidence-based units was 4% (SD = 0.06) at pretest and 3% (SD = 0.05) at posttest, a non-significant change.
For the control condition, the percentage of claim-congruent functional evidence-based units was 33% (SD = 0.16) at pretest and 21% (SD = 0.13) at posttest, a significant decrease, \( F(1, 49) = 19.790, p < 0.0005 \), Partial \( \eta^2 = .288 \). The percentage of claim-incongruent functional evidence-based units was 2% (SD = 0.05) at pretest and 3% (SD = 0.05) at posttest, a non-significant change. In conclusion, when we took into account changes in essay length from pretest to posttest, only the ER+AP condition showed enhanced performance from pretest to posttest.

Next we examine *However*-related units at posttest. Again, we used negative binomial regression with condition as a predictor factor and pretest as a covariate. For evidence-based *However* units (Figure 11), at posttest, the mean number was 1.56 (SD = 1.19) units for the ER+AP condition, 0.56 (SD = 0.70) units for the AP condition, and 0.44 (SD = 0.70) for the Control condition, a significant difference across the three conditions, Wald \( \chi^2(2) = 27.925, p < 0.0005 \). The ER+AP condition had an expected 2.892 (95% CI, 1.600 to 5.228) times more evidence-based *However* units than the AP condition, a significant result, Wald \( \chi^2(1) = 12.358, p < 0.0005 \). The AP condition and Control condition did not differ significantly in the mean number of evidence-based *However* units, Wald \( \chi^2(1) = 0.412, p = 0.521 \).
Figure 11. Mean number of evidence-based *However* units by condition at pretest and posttest

![Graph showing the mean number of evidence-based *However* units by condition at pretest and posttest. The graph displays three conditions: ER+AP, AP, and Control. The ER+AP condition shows a significant increase in units from pretest to posttest, while the AP and Control conditions show less increase.](image)

For full evidence-based *However* units (Figure 12), at posttest, the mean number was 0.56 (SD = 0.80) units for the ER+AP condition, 0.15 (SD = 0.36) units for the AP condition, and 0.20 (SD = 0.45) for the Control condition, a significant difference across the three conditions, Wald $\chi^2(2) = 9.825, p = 0.007$. The ER+AP condition had an expected 3.980 (95% CI, 1.314 to 12.053) times more full evidence-based *However* units than the AP condition, a significant result, Wald $\chi^2(1) = 5.968, p = 0.015$. The AP condition and Control condition did not differ significantly in the mean number of full evidence-based *However* units, Wald $\chi^2(1) = 0.302, p = 0.583$. 
McNemar’s test was performed to examine change in the percentage of students who ever included any of the above categories from pretest to posttest within each of the three conditions. McNemar’s test was selected for statistical analysis because it is often used in pretest-posttest design with a dichotomous dependent variable.

As shown in Table 8, for percentage of students who ever included a functional evidence-based unit, there was no significant change from pretest to posttest for any of the three conditions. For percentage of students who ever included a claim-congruent functional evidence-based unit, there was also no significant change from pretest to posttest for any of the three conditions. For percentage of students who ever included a claim-incongruent functional evidence-based unit, the proportion increased from 23% to 70% for the ER+AP condition, a statistically significant increase, \( p = .002 \). The change from pretest to posttest for the other two conditions did not reach statistical significance.
Table 8. Percentage of students who ever made a certain type evidence-based claim by condition at pretest and posttest

<table>
<thead>
<tr>
<th></th>
<th>Pretest</th>
<th></th>
<th></th>
<th>Posttest</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ER+AP</td>
<td>AP</td>
<td>Control</td>
<td>ER+AP</td>
<td>AP</td>
<td>Control</td>
</tr>
<tr>
<td>Claim-congruent</td>
<td>100%</td>
<td>93%</td>
<td>96%</td>
<td>100%</td>
<td>93%</td>
<td>88%</td>
</tr>
<tr>
<td>functional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Claim-incongruent</td>
<td>23%</td>
<td>37%</td>
<td>20%</td>
<td>70%</td>
<td>41%</td>
<td>30%</td>
</tr>
<tr>
<td>functional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td>100%</td>
<td>96%</td>
<td>96%</td>
<td>100%</td>
<td>96%</td>
<td>90%</td>
</tr>
</tbody>
</table>

As shown in Table 9, for percentage of students who ever included an evidence-based

_However_ unit, we observed an increase from 27% to 74% for the ER+AP condition, a

statistically significant increase, _p_ = 0.002. The change from pretest to posttest for evidence-based _However_ units did not reach statistical significance for the other two conditions.

For percentage of students who ever included a full evidence-based _However_ unit, we observed an increase from 8% to 41% for the ER+AP condition, a statistically significant increase, _p_ = .012. The change from pretest to posttest for full evidence-based _However_ units did not reach statistical significance for the other two conditions. Overall, individual-level analysis indicated that the RP condition was the only condition that demonstrated significant improvement from pretest to posttest in the more challenging task of addressing claim-incongruent evidence showing full evidence-based _However_ units.
Table 9. Percentage of students who ever made an evidence-based *However* units and full evidence-based *However* units by condition at pretest and posttest

<table>
<thead>
<tr>
<th></th>
<th>Pretest</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ER+AP condition</td>
<td>AP condition</td>
</tr>
<tr>
<td>Evidence-based <em>However</em></td>
<td>27%</td>
<td>33%</td>
</tr>
<tr>
<td>Full evidence-based <em>However</em></td>
<td>8%</td>
<td>15%</td>
</tr>
</tbody>
</table>

Qualitative portrayal of change in essays

Having presented quantitative data to demonstrate change over time for the two Intervention conditions, particularly the ER+AP condition, and lack of change for the Control condition, we undertake here to present qualitative data to portray the nature of progression over time. To accomplish this goal, we translated the following uncorrected essays from Chinese to English. We inserted the numbering of idea units to assist in analyzing these essays.

We focus in this section on two students’ essays: Student A from the Control condition and Student B from the ER+AP condition. Both Student A and Student B were representative of their conditions. As will be analyzed in more detail, Student A’s pretest and posttest essays demonstrated minimal change over time. Student B’s essays, on the other hand, demonstrated increasingly sophisticated integration of advanced argumentative elements such as evidence-based *However* units. Student B also progressed in referring to an increasingly wider array of evidence both consistent and inconsistent with his position.

Student A pretest essay  Position: Juvenile court

<table>
<thead>
<tr>
<th>Idea unit</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adult court and juvenile court exert different influence on a teenager.</td>
</tr>
<tr>
<td>2</td>
<td>Based on the list of information I received regarding juvenile court and adult</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3</td>
<td>For instance, the record of a juvenile trial will not be made public. If the record is made public, it will negatively influence the future development of that teen.</td>
</tr>
<tr>
<td>4</td>
<td>So we should take the future development of a teenager into consideration and minimize the negative psychological influence.</td>
</tr>
<tr>
<td>5</td>
<td>Adults also think more carefully before they act, but the crime a teenager commits will have lasting influence on him or her.</td>
</tr>
<tr>
<td>6</td>
<td>In order to minimize the chance that the teen will commit a crime again, we need to have a teen tried in juvenile court.</td>
</tr>
<tr>
<td>7</td>
<td>Also, I think it is a good idea to separate teens from adults. The evidence says that teens in adult jails are 50% more likely to be attacked by another inmate. So we should put teens in a different environment.</td>
</tr>
</tbody>
</table>

In Student A’s pretest essay, we clearly see a dual focus in which arguments both supporting her position and weakening the opposing position were marshaled and applied. A particularly noticeable aspect was Student A’s ability to make an extrapolation from existing evidence that originally supported her position in order to weaken the opposing position. For instance, in Idea unit 3, she made an extrapolation from evidence that records in juvenile court are sealed upon release; by hypothesizing a scenario in which “if the record is made public…” she now turned the evidence around to weaken the adult court position.

We also see a hint of an integrative structure in Idea unit 2, “I think the positives of a juvenile court outweigh the negatives,” which suggested that Student A might have taken into consideration both the positives and negatives of her position. However, she failed to make the integration explicit and the rest of the essay did not mention any of the negative aspect of her position. We also noticed some non-functional use of evidence in Idea unit 5, as the evidence cited that “adult think more carefully before they act” was not explicitly or successfully connected to the claim in Idea unit 6. Overall, Student A’s pretest essay did not integrate evidence inconsistent with her position. The However structure was also entirely missing in her pretest essay.
Hello, as an average student in middle school, I am really against the idea that teenagers who have committed serious crimes should be tried in adult court.

This is because residents in China are mostly middle-aged adults, and teens only take up a small proportion.

The information says that teens in adult jails are 50% more likely to be attacked by another inmate and twice as likely by prison staff. This shows how cruel the adult system is.

Besides, the prefrontal cortex responsible for good judgment is not fully developed until 25. So if a teenager is tried and served in an adult jail, it will distort his or her thinking and value system.

This also explains why it is common for teenagers to have a high rate of recidivism.

But if a teenager is tried in a juvenile court, there are staff specifically trained to deal with young people, which means once the teen leave the juvenile prison he will maintain confidence in life and will contribute to our country’s development.

Let me reiterate, teens should not be tried in an adult court!

At posttest, we see that Student A sticks to her pretest position that teenagers should be tried in a juvenile court. Student A retained her dual-focus structure at posttest, and she supported her position in Idea unit 6 and weakened the opposing position in Idea units 3, 4, 5.

We noted that she again made an extrapolation from existing evidence to weaken the opposing side. For instance, in Idea unit 4, having cited the evidence that “the prefrontal cortex…is not fully developed until 25,” she immediately inferred that if a teenager was tried in adult jail, “it will distort his or her thinking and value system,” which weakened the opposing side. Another noticeable aspect of her posttest essay was that she made a causal inference in Idea units 4 and 5, using possible “distortion of teen’s thinking and value system” in adult jail in unit 4 as a causal factor to explain another piece of evidence regarding teenager’s “high rate of recidivism” in unit 5. Again, Student A’s posttest essay did not integrate evidence inconsistent with her position and did not construct the advanced However unit.
Student B pretest essay       Position: Juvenile court

<table>
<thead>
<tr>
<th>Idea unit</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I favor the position that they should be tried in a juvenile court.</td>
</tr>
<tr>
<td>2</td>
<td>In 2006 in China, 70% of juveniles who committed serious crimes are subjected to sentences for less than 3 years in juvenile court.</td>
</tr>
<tr>
<td>3</td>
<td>A human brain reaches maturity at age 25, in other words, the decision-making and thinking ability of a teenager is not mature yet, and they do not have good control of themselves so they should be tried in a teen court.</td>
</tr>
<tr>
<td>4</td>
<td>The information list also says that teens in adult jails are 50% more likely to be attacked by another inmate and twice as likely by prison staff, compared to adult prisoners.</td>
</tr>
<tr>
<td>5</td>
<td>Also, the judges and staff in a juvenile system are specially trained to deal with young people who have committed crimes.</td>
</tr>
<tr>
<td>6</td>
<td>In comparison, adult court does not have these well-trained staff and that’s why it is not suitable for teenagers.</td>
</tr>
</tbody>
</table>

Student B in his pretest essay also took on a dual focus by supporting his position in Idea unit 3 and weakening the opposing position in Idea unit 6. We observed non-functional use of evidence in Idea units 2 and 4, as they were verbatim copies of provided evidence and Student B failed to make any attempt to connect them to any claim. Similar to Student A, Student B extrapolated from evidence that supported his position, “judges and staff in juvenile system are specifically trained to deal with young people,” to weaken the opposing position in Idea unit 6, “adult court does not have these well-trained staff.” Similar to Student A in her pretest essay, Student B did not address evidence inconsistent with his position and did not integrate the However structure into his essay.

Student B Topic 1       Position: Part-time job for high school students

<table>
<thead>
<tr>
<th>Idea unit</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>After spending a month discussing this topic, I now have a much deeper understanding of the topic.</td>
</tr>
<tr>
<td>2</td>
<td>I continue to support the position that high school students should take on a part-time job.</td>
</tr>
<tr>
<td>3</td>
<td>I think that without work experience a person can’t function in a society in the future.</td>
</tr>
<tr>
<td>4</td>
<td>A study in America shows that high school students who have work</td>
</tr>
</tbody>
</table>
experience have clearer understanding of their career development.

5 And of course, this is not the only reason why I think high school students should work.

6 First of all, I think if you only focus on study, you wouldn’t even know how to take care of yourself in real life.

9 Without any experience, how can you settle down in society in the future?

10 Our society does not need a bookworm; we need a well-rounded person.

11 It is true that many people will say that high school curriculum is heavy and there is no time left to do a part-time job.*

12 However, I don’t think the problem here is real.

13 You can think of it as a test of your ability to allocate your precious time wisely.

14 If you can’t even handle this, what can you do once you enter society and face even more tasks? Are you going to go crazy?

15 If you really can’t do a part-time job during the school year, you can do it during your winter or summer break.

*Note. Sequences of However units appear in bold

Student B opened his Topic 1 essay by making a metacognitive statement that after discussing the topic, he “now has a much deeper understanding of the topic.” In the rest of his Topic 1 essay, we observed a progression from a dualistic focus to more integrative thinking that considered criticism of his position. For instance, in Idea unit 11, he acknowledged a criticism from a piece of evidence that “high school curriculum is heavy.” Subsequently, he successfully countered this criticism in units 12, 13, 14. He also offered an alternative solution, which was to do a part-time job “during your winter or summer break” in unit 15.

Student B Topic 2 Position: Allow animal research

<table>
<thead>
<tr>
<th>Idea unit</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I believe that animals can be used in medical research.</td>
</tr>
<tr>
<td>2</td>
<td>The primary reason why we can use animals is that organs of animals and humans are very similar.</td>
</tr>
<tr>
<td>3</td>
<td>In order to advance the medicine of humans, we have to sacrifice some animals for medical research.</td>
</tr>
<tr>
<td>4</td>
<td>The other side says that animal research is too cruel.</td>
</tr>
<tr>
<td>5</td>
<td>However, what I want to say is that research shows that a lot of drugs that help to treat diseased animals come from research on animals. This means that animal research isn’t all bad.*</td>
</tr>
<tr>
<td>6</td>
<td>In order to protect animals, we need to address the root problem.</td>
</tr>
<tr>
<td>7</td>
<td>The other side says that not all drugs that pass the animal test are successful on humans.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>8</td>
<td>While this is true, I think even though the success rate isn’t that high, one success still means a lot.</td>
</tr>
<tr>
<td>9</td>
<td>The medicine of human makes a huge leap forward with each successful drug that comes out.</td>
</tr>
<tr>
<td>10</td>
<td>If we agree with the opposing side’s opinion, are we going to stop the advancement in medicine just to protect animals?</td>
</tr>
<tr>
<td>11</td>
<td>The other side also says that we can use models of human bodies for research.</td>
</tr>
<tr>
<td>12</td>
<td>However, a lot of research shows that a lot of studies of living bodies are so complicated and uncertain that they could only be carried out with animals. For example, studies in gene engineering test how to modify the organs of animals so they can be transplanted to humans. This type of medical research is so complicated and a simple model of the human organs is far from enough.</td>
</tr>
<tr>
<td>13</td>
<td>Animal testing has led to treatments and cures for many human diseases such as AIDS, diabetes and hepatitis.</td>
</tr>
<tr>
<td>14</td>
<td>Imagine the number of people that will be saved by the drugs from animal testing when there is an outbreak of these diseases in the future.</td>
</tr>
<tr>
<td>15</td>
<td>Regulations exist that require that scientists use as few animals as possible to conduct their research.</td>
</tr>
<tr>
<td>16</td>
<td>Also, most animals used in medical testing are caged animals bred specifically for this purpose.</td>
</tr>
<tr>
<td>17</td>
<td>So even if we release these animals to nature they probably couldn’t even survive, so we don’t even have to worry about the damage to the ecosystem.</td>
</tr>
<tr>
<td>19</td>
<td>Overall, I think animals can be used in research.</td>
</tr>
</tbody>
</table>

*Note. Sequences of *However* units appear in bold

Student B’s Topic 2 essay made further progress, as the essay became solidly integrative. Acknowledgements that supported opposing side and weakened own side were made in Idea units 4, 7, 11, with Idea units 7 and 11 coded as evidence-based. All three units were successfully countered in subsequent statements, making all of them evidence-based *However* units. For instance, the evidence that “animal research benefits animals as well” in Idea unit 5 successfully countered the weaken-own statement that “animal research is too cruel” in Idea unit 4. In this essay, we could clearly observe the classic argument → counterargument → rebuttal structure, which indicated skilled argumentive writing.
I am supporting the position that kidney sales should be allowed. First of all, in China in 2012, for every 150 people waiting for a donated kidney, 149 die while waiting. This clearly means that donated kidneys are not enough for the need of our society so we need to legalize selling to meet the need.

Even though the opposing side has cited that in 2016, over 100,000 people registered to be organ donors, a threefold increase over 2015, the reality is that in China, there is about a million people on waiting list to receive kidney transplants. However, only about 10,000 kidney transplant surgeries are carried out each year.*

So many people die because there is no suitable kidney to be transplanted. The opposing side also gets concerned because the price of a kidney in a legal market might be as high as 300,000 to 40,000 RMB, which exceeds the annual income of an average family.

However, if you stay on dialysis to stay alive you have to pay 499,220 RMB a year, which also exceeds the annual income of a family.

Also, your life expectancy is less than 8 years on dialysis, so you are better off just buying a kidney. From the perspective of a person who wants to sell a kidney, almost always they are very poor and have few ways to earn money. For instance, some kidney sellers in Iran need the money to pay off debts.

Well, if a grown-up wants to sell a kidney for money, why should we stop him from doing it? The other side says donation will give equal opportunity of life for all people, poor or rich.

However, rich people will always find a way to secure an illegal kidney. If our country legalizes kidney selling, I don’t think the leaders in our country will leave the market to itself. The government will set up laws to control the price of the kidney market.

The other side says that if the market is legalized, rich people will buy in a lot of kidneys and sell them out at a very high price.

However, just like the anti-monopoly law in China right now, our country won't allow anyone or any company to monopolize the kidney market.

The statistics in America say that in 2010, 1/5 of the kidneys come from the black market. We should prevent this situation from happening by allowing kidney sales in a transparent market.

Our current policy regarding kidney has many problems, so we need to modify or change it completely.

*Note. Sequences of However units appearing in bold
Student B’s Topic 3 essay maintained the advanced integrative structure and we identified *However* units in Idea units 4, 7, 12 and 15. Similar to Student B’s Topic 2 essay, most of these *However* units were evidence-based. It should be noted that in unit 15, Student B introduced a piece of evidence from his own knowledge source regarding “anti-monopoly law in China.” We are unsure where he acquired this piece of information but he successfully connected it to his argument that “China won’t allow anyone or any company to monopolize the kidney market,” countering the criticism from the opposing side that rich people might hoard kidneys if selling is allowed.

Overall, having examined Student B’s three intervention topic essays, we are confident that the dialogic structure of argumentation with the opposing side had clearly made its way into this student’s essays (e.g., “The other side says…”). The continuing experience of dialogue with the opposing side helped the writer to develop the *However* structure, which enabled him to first introduce arguments and/or evidence inconsistent with his position, followed by rebuttal that countered them.

<table>
<thead>
<tr>
<th>Student B Posttest essay</th>
<th>Position: Juvenile court</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idea unit</td>
<td>Content</td>
</tr>
<tr>
<td>1</td>
<td>The judges and staff in a juvenile system are specially trained to deal with young people who have committed crimes, which means that teenagers who are tried in a teen court would not be placed at a huge disadvantage.</td>
</tr>
<tr>
<td>2</td>
<td>Also, the prefrontal cortex that is responsible for abstract thinking and the ability to exercise good judgment is not fully developed until one is 25.</td>
</tr>
<tr>
<td>3</td>
<td>Another piece of information says that teenagers do not have the same level of self-control as adults.</td>
</tr>
<tr>
<td>4</td>
<td>As a result, I don’t think it is suitable for teenagers to be tried in an adult court.</td>
</tr>
<tr>
<td>5</td>
<td>Punishments tend to be less severe and sentences shorter in juvenile court. In 2006 in China, 70% of juveniles who committed serious crimes are subjected to sentences for less than 3 years in juvenile court</td>
</tr>
<tr>
<td>6</td>
<td>I think this is a great idea because teenagers are young so we need to give them more years outside of prison to correct their mistakes.</td>
</tr>
<tr>
<td>7</td>
<td><strong>But here comes the problem, adult jails provide job training for inmates</strong></td>
</tr>
</tbody>
</table>
by subjecting them to labor reform, which means adults wouldn’t be jobless when leaving prison.

8 However, there is a risk in adult prison too as teens in adult jails are 50% more likely to be attacked by another inmate and twice as likely by prison staff. *

9 And this will cause severe damage to a teenager.

10 Also, the record in juvenile court would not be made public and is sealed upon release, this helps teenagers “save face” in future life.

11 However, if we help them “save face” and make them feel like there is no real consequence for their bad behavior, the rate of repeat crime becomes 90% over 10 years.

12 What we can do is to subject teenagers in juvenile jail to better education and reform efforts, which might lower the rate of recidivism in the future.

13 No matter what, if we can prevent teenagers from committing crimes, we would save ourselves from all these trouble.

*Note. Sequences of However units appear in bold

Student B’s posttest essay displayed transfer of skills in making integrative statements from intervention topics to a non-intervention topic. We identified two However statements in Idea units 8 and 11. Idea Units 7 and 8 form a full evidence-based However unit, as they both referenced evidence when discussing the affordance and risk of adult prison. Idea units 10 and 11 formed a typical However unit in that the writer first discussed the benefit of juvenile court referring to the evidence that “record in juvenile court would not be made public.” He then criticized the evidence by citing the high rate of repeat crime among teenagers. In Unit 12, the writer attempted to offer a resolution, which was to “subject teenagers in juvenile court to better education and reform efforts.” The apparent integrative thinking exhibited in Student B’s posttest essay was particularly encouraging, in that the writer was able to introduce criticism of his position, followed by successful rebuttal to these counterarguments, now in the absence of participation in any form of dialogic argumentation before writing.

Analyses of evidence reflection sheets
The analyses carried out above point to some overall change among students in the ER+AP condition in terms of enhanced disposition and competence to address evidence both congruent and incongruent with their favored position. What was the underlying mechanism driving these changes? While both the ER+AP condition and AP groups engaged in the dialogic argument curriculum, we believe it was the added evidence reflection component that gave students in the ER+AP condition further advantage. An analysis of ER+AP students’ answers to the evidence reflection sheets aims at providing observations, more of a speculative nature, of the underlying cognitive changes that may have contributed to the greater gains observed in the performance of students in the ER+AP condition.

During each of the dialogue sessions, a piece of evidence was made available to pairs and they were verbally prompted to address it when dialoguing with the opposing side. An added procedure in the ER+AP condition was that pairs were also asked to jointly complete evidence reflection prompts during these dialogue sessions. The evidence reflection sheets consisted of Part A and Part B. Part A, completed prior to the start of dialogue, asked pairs to anticipate the potential use of that piece of evidence in dialogue, with particular focus on how the evidence could be countered and how the opposing side might use it if the evidence was made available to them. Part B, completed toward the end of the dialogue, asked pairs to evaluate their actual use of evidence in the dialogue, with particular emphasis on whether they were satisfied or dissatisfied with the use, and why.

We focused on analyzing students’ answers in dialogue 1 (M+ evidence), dialogue 2 (O-evidence), dialogue 3 (O+ evidence) and dialogue 4 (M- evidence) of each of the three topics. We collapsed data for dialogue 1 (M+ evidence) and 2 (O- evidence) as the evidence was claim-congruent; we also collapsed data for dialogue 3 (O+ evidence) and 4 (M- evidence) as the
evidence was claim-incongruent. As a result, there were 28 pair responses for claim-congruent evidence and 28 pair responses for claim-incongruent evidence for each of the three topics.

Anticipation of evidence use (Evidence Reflection Sheet – Part A)

Question one in Evidence Reflection Sheet – Part A asked pairs to recall whether the evidence had been mentioned by the opposing side in previous dialogues. Question two asked pairs to judge whether the evidence helped their side or the opposing side. Data showed that pairs were mostly accurate in their response to these two questions for all three topics, suggesting that pairs have engaged in meta-level (second-level) processing of the evidence, the first level being comprehension of the information represented in the evidence.

A notable change over time emerged in pairs’ answers to Question four, “How might someone who disagree with you on the issue use this evidence?” The change was most noticeable for claim-congruent evidence, which represented claim-incongruent evidence for the opposing side. During Topic 1, 22 pair responses (79%) said that the other side would not use it because the evidence would not help them. During Topic 2, 14 responses (50%) said the other side would not use it. During Topic 3, only 12 responses (43%) said this.

Given that the majority of pairs answered that the other side would use the claim-incongruent evidence in Topic 3, we further analyzed their answers regarding how the other side might use it. Of the 12 responses that said the other side would use it, 7 provided a specific counterargument to this piece of evidence, 3 said the opposing side could use another piece of evidence to counter it but did not specify which piece of evidence, and 2 said that the other side could search for more information that would contradict this piece of evidence.
Evaluation of evidence use (Evidence Reflection Sheet – Part B)

Question one in Evidence Reflection Sheet – Part B asked participants whether they managed to address the evidence during dialogue, and, if not, why. For claim-congruent evidence, the majority of pairs said they used it for all three topics. But for claim-incongruent evidence, the majority of pairs (25 responses, 89%) at Topic 1 indicated that they did not address it because the evidence would not help them. This number decreased to 18 (64%) at Topic 2 and further decreased to 14 (50%) at Topic 3.

Question two in Part B asked participants to evaluate their use by judging their level of satisfaction and by providing reasons for their satisfaction (or dissatisfaction). Here we observed another pattern over time. Pairs were initially highly satisfied with their use of claim-congruent evidence, but became less satisfied as time passed. On the other hand, pairs were highly dissatisfied with their use of claim-incongruent evidence, but became more satisfied as time passed. Table 10 presents the percentage of pairs who said they used evidence and those who were satisfied with their use.

Analysis of pairs’ answers to why they were satisfied or dissatisfied yielded interesting findings as well. For claim-congruent evidence, a common reason for satisfaction was that “the evidence provided concrete support for my reason so it makes it difficult for the opposing side to counter it.” Over time, as students grew increasingly dissatisfied, a common reason for dissatisfaction was “our evidence was easily countered by our opponents,” indicating that students had begun to anticipate counters to their evidence. For claim-incongruent evidence, a common reason for dissatisfaction at the beginning was “the evidence does not help us, so it actually makes us weak.” Over time, a common reason for their satisfaction was “we must be
really strong if we are able to counter a piece of evidence that does not help us.” They thus recognized a need to address all evidence, not just supporting evidence.

Table 10. Percentage analysis for evidence reflection sheet – part B

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Evidence type</th>
<th>Percentage of pairs who said they used the evidence</th>
<th>Percentage of pairs who used the evidence said they were satisfied with the use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>Claim-congruent</td>
<td>89%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Claim-incongruent</td>
<td>11%</td>
<td>0%</td>
</tr>
<tr>
<td>Topic 2</td>
<td>Claim-congruent</td>
<td>93%</td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td>Claim-incongruent</td>
<td>36%</td>
<td>40%</td>
</tr>
<tr>
<td>Topic 3</td>
<td>Claim-congruent</td>
<td>100%</td>
<td>57%</td>
</tr>
<tr>
<td></td>
<td>Claim-incongruent</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>
DISCUSSION

The present study examined whether, in the context of dialogic argumentation, joint reflective activities regarding evidence use would lead to greater gains in middle school students’ ability to coordinate claims with evidence in individual written arguments. Overall, the intervention examined in the present work proved successful. The added component of engaging student dyads in joint meta-level reflection had substantial effects on developing students’ skills in addressing evidence both congruent and incongruent with one’s own claims. The results ruled out a competing hypothesis that since students in the Argument Practice (AP) condition did not devote time to the two evidence reflection sheets that distinguished the experimental (ER+AP) condition, they would show superior performance given that they had more practice in evidence-based dialogic argumentation. In fact, the AP group showed inferior performance, despite this extra practice. In what follows, I first summarize the major findings of the study, followed by discussion of implications, limitations, and future directions.

Principal findings

A developmental pattern emerged across time during students’ engagement with the three intervention topics. During Topic 1, in addressing evidence congruent with their claim (either supports own position or weakens opposing position), the two groups showed parallel performance, with 89% of ER+AP students and 92% of AP students ever making this type of evidence-based claim. In addressing evidence incongruent with their claim (either supports opposing side or weakens own side), the two groups showed a remarkable difference, with 62% of ER+AP students and 19% of AP students ever making this type of evidence-based claim. At Topic 1, the average frequency of claim-incongruent evidence-based claims was 0.77 for the
ER+AP condition and 0.19 for the AP condition, frequencies well below one unit. Thus there is still much room for further development for both conditions.

Anecdotal evidence suggested that Topic 2 was a topic for which students showed particularly high level of motivation and engagement. Topic 2 asked students to argue about whether animals could be used in medical research. Students were reported to have spontaneously consulted their biology teacher on the issue, both in and out of their biology class. Both conditions showed remarkable progress from Topic 1 to Topic 2 and there wasn’t significant difference across the two conditions in terms of the mean frequency of the two types of evidence-based claims. However, when taking essay length (idea units) into account, the ER+AP condition outperformed the AP condition.

In the last intervention topic, decisive advantage of the ER+AP condition over the AP condition appeared. Although 96% of students from both conditions made claim-congruent evidence-based claims at least once, the ER+AP condition showed a mean frequency of 4.62 units, significantly higher than that of 3.04 for the AP condition. For claim-incongruent evidence-based claims, 81% of ER+AP students, contrasted with 33% of AP students, ever made use of them. For the ER+AP condition, the mean frequency of making this challenging type of evidence-based claim reached 1.65 units at Topic 3, a significant gain from the 0.77 units at Topic 1.

In measuring students’ construction of *However* statements that connect two conflicting statements, we saw that while the ER+AP group was significantly more successful in referring to evidence in making *However* statements as early as Topic 1, it was only by Topic 3 that they became significantly more successful in connecting two conflicting pieces of evidence, a skill that we expected would take longer to develop.
Overall, analysis of intervention topic essays suggested that while adding a reflection component led to immediate advantage in performance at the end of Topic 1, it was only by the end of the last topic that a decisive superiority in addressing evidence both congruent and incongruent with one’s claims was observed. By the last topic, when students introduced a piece of claim-incongruent evidence, it was often followed by another piece of claim-congruent evidence. We coded these cases as full evidence-based *However* units, and we viewed them as constituting a successful argumentative strategy sequence that restored the strength of one’s claim. The sequence also resembled the argument-counterargument-rebuttal structure suggestive of advanced argumentative writing (Leitao, 2000; Nussbaum & Schraw, 2007).

For the transfer topic, students who were exposed to joint reflective training were better able to transfer their skills of constructing evidence-based claims from intervention topics to a new topic. While both intervention conditions outperformed the no-treatment control condition in constructing claim-congruent evidence-based claims, only the evidence reflection group showed a clear advantage in the more challenging task of voicing claim-incongruent evidence-based claims. In fact, 70% of ER+AP students included a claim-incongruent evidence-based claim in their writing at least once, a significant increase from the 23% during pretest, and only slightly less than the 81% during the last intervention topic. In contrast, only 41% of AP students and 30% of control students ever included any claim-incongruent evidence-based claims in their posttest essays.

**Educational and theoretical implications**

The present study is consistent with prior studies (e.g., Felton, 2004; Nussbaum et al., 2018; Pressley & Ghatala, 1990) that have shown that engaging students in extended practice of
a certain skill alone does not necessarily lead to superior performance, compared to interventions that combine practice with metacognitive training. In the present study, dyads in the ER+AP condition spent on average 10-12 minutes to complete the evidence reflection sheets, during each of the 40-minute dialogue sessions. During this joint reflective activity, students were instructed to discuss fully with their dyadic partner, reach an agreement, and then write down their answers.

For each topic, students worked with their dyadic partner for five sessions. Previous research indicates that working with a same partner for extended sessions helped partners develop a collaborative relationship, which enabled them to better profit from the regulation of one another’s thinking and behavior (Zillmer & Kuhn, 2018). Previous research also indicated that symmetric peers with equal abilities and social standing are able to regulate one another, adopting the roles of teacher and learner interchangeably (Zillmer & Kuhn, 2018; Forman & Cazden, 1985), thus extending Vygotsky’s portrayal of the Zone of Proximal Development in which a more capable individual guides a less capable one.

It should be noted that the focus of the present study was not on examining the exchange within dyads to reveal the nature of their collaboration. Rather, the focus was on analyzing the products of such collaboration, in the form of the written responses students provided in completing the evidence reflection sheets. Another major focus was whether this metastrategic training would lead to higher strategic performance (Kuhn & Pearsall, 1998) as measured by the frequency of evidence-based claims in students’ individual written arguments.

We observed a topic content effect in that when students’ interest and engagement with a topic was high, such as in Topic 2 in the present study, both conditions showed parallel high level of performance. However, when a new topic with more challenging content was introduced, such as in Topic 3 regarding legalization of kidney sales, only students in the
evidence reflection condition were able to maintain their high level of performance and even made further progress. Students who did not engage in evidence reflection showed diminished performance. Engagement in joint meta-level reflection thus seemed compensatory when students’ interest and engagement in a topic was relatively low, a finding with important implications.

When teachers adopt the argument curriculum used here, or dialogic teaching that emphasizes student discussion in general, there will always be variation in student interest in the topic or content under discussion. Rather than avoiding knowledge-dense topics that students struggle with initially, teachers could consider incorporating joint reflective exercises that promote peer-to-peer discussion and collaboration that would potentially lead to high level of performance.

Earlier we discussed findings of positive transfer from intervention topics to a new topic for ER+AP students. The transfer here is near transfer since the writing task is similar to that during the intervention, just on a new topic. In fact, in addition to these near transfer effects, a small amount of studies on argumentation-based learning have identified that structured classroom argumentation could also lead to far transfer, including increases on standardized test scores years after the intervention, transfer to other disciplines, and rises in fluid intelligence (Nussbaum & Asterhan, 2016). Several theories have been proposed to explain these positive transfer effects. Adey and Shayer (2015) invoked a Piagetian framework, claiming that dialogue-based instruction helps to develop formal operational thought, when explaining far transfer effects of cognitive skills from Science to English Language Arts. Nussbaum and Edward (2011) referred to neo-Piagetian notions, which suggest that advances in argumentive writing are driven
by a process of differentiation and integration, when explaining transfer of argument-counterargument integration skills to a novel problem.

To account for findings of positive transfer in the present study, we refer to previous research proposing trainings that target meta-level development to enable transfer of skills to a new task (Kuhn, 2000; Felton, 2004). By adding a joint evidence reflection component into dialogue sessions, the advantage of the original Q&A method was amplified (Kuhn & Moore, 2015; Hemberger et al., 2017), since the combination of practice and reflection in the ER+AP condition better facilitates the development of argument schema (Bereiter & Scardamalia, 1982; Reznitskaya & Anderson, 2002).

We propose that development of argument schema regarding evidence use enables transfer of skills from familiar topics to a novel topic. Engagement in reflection during dialogic argumentation helps direct students’ attention to relevant evidence, promote comprehension of the challenging claim-incongruent evidence, and facilitate coordination of claim-incongruent and claim-congruent evidence using the advanced However structure. The similarity between the intervention task and the transfer task, including the Q&A format of evidence provision, activates the argumentation schema (Reznitskaya & Anderson, 2002), which enables transfer across topics.

We also propose transfer from collaborative reflection to individual reflection, based on the sociocultural account of the internalization of higher mental processes from social plane to individual plane (Vygotsky, 1978; Rogoff, 1990). We assume that students who engaged in repeated reflection on evidence in a dyadic setting would engage in similar reflective self-regulatory activities when they wrote individually on the same topic, asking themselves questions such as “Does the evidence support my side or other side?” “How might I counter this
piece of evidence supporting the other side?” and “Am I satisfied with my use of the evidence and how might I improve it?” Thus, we believe that engagement in joint meta-level exercises help students develop a reflective “habit of mind,” which was then carried over into an individual setting.

However, not all skills successfully transferred to a new task. Comparing the performance of the ER+AP condition between the last intervention topic (Topic 3) and transfer topic, we notice a decrease in the mean frequency of evidence-based However statements (from 2.81 units to 1.56 units) and full evidence-based However statements (from 1.65 units to 0.56 units). The percentage of students ever making a full evidence-based However statement decreased from 65% at Topic 3 to 41% at transfer topic. Thus, even though meta-level training supported transfer of skills of constructing evidence-based claims to an unfamiliar topic, students’ lack of engagement in dialogic argumentation on the topic diminished the strong presence of the dialogic structure (measured by the frequency of evidence-based However statements) observed in their individual writing in the last intervention topic.

The findings supported our claim that dialogue and deep engagement with the topic are the two most critical dimensions of our dialogic approach, since they made the voice of an opposing side prominent to students, so much so that even when students write individually, their writing frequently reflects the dialogic structure. “Others might say…However…” became a frequent structure in essays, as illustrated in Student B’s writings provided in the Results section. When dialogic argumentation is removed, as is the case during the transfer task, we observed a decrease in the frequency with which students integrate two pieces of conflicting evidence, which indicates diminished two-sided thinking when arguing with evidence.
Nevertheless, the superior performance of ER+AP students in the last intervention topic and transfer topic warranted a closer examination at the responses they provided to the evidence reflection sheets, which would shed light on the mechanism that accelerated their development. Previous research indicated that using anomalous data to promote students’ theory change is not an automatic process. Rather, the first step is to bring individuals to realize the existence of a meaningful conflict between their own theories and the evidence conflicting with them (Mason, 2000; Chinn & Brewer, 1993). Thus, one of the first few questions in the evidence reflection sheet asked students whether the evidence they received helped to support their side or their opponents’ side. The finding that students were nearly 100% accurate in their answers to this question suggested that we successfully brought students’ attention to consider whether there was a conflict between the evidence and their position.

In anticipating their potential use of the evidence, students were asked to generate possible counters to the evidence. For evidence congruent with their claim, they were expected to anticipate how the opposing side would counter the evidence. For evidence incongruent with their claim, they were expected to anticipate how they could counter the evidence. Through these reflective practices, students developed not only the inclination but also the skills to consider how a piece of evidence could be undermined by counterargument. They were also prompted to consider how the opposing side would use the evidence if they received it. Recall that in our design of distributing evidence to students, the two sides received the same set of evidence but in different order. As a result, students might well already be aware that the evidence they received would be made available to the opposing side, which gave anticipating how the other side would use it a practical purpose.
In evaluating their actual use of the evidence, if students did not address the evidence in their dialogue, they were asked to reflect on the reasons why they did not do so. If students addressed the evidence, they were asked to rate their satisfaction level and explain why they were satisfied, which prompted dyads to discuss and generate evaluative criteria for evidence use in argumentation. Here, an interesting developmental trend was observed. Dyads were highly satisfied with their use of claim-congruent evidence at Topic 1 (100% of responses) but grew increasingly dissatisfied over time (57% of responses). At the same time, they were highly dissatisfied with their use of claim-incongruent evidence at Topic 1 (0% of responses), but grew increasingly satisfied over time (50% of responses). The trend suggested that dyads became increasingly critical of their use of claim-congruent evidence, possibly because their opponents countered their evidence-based claim, which prompted them to realize the inadequacy in their initial evidence-based claim. At the same time, they became more accepting of their use of claim-incongruent evidence, and a common reason they provided for their satisfaction was “we are really strong if we are able to counter a piece of evidence that did not help us.”

The findings yielded support for our design to facilitate students gaining information about a topic by providing them a piece of evidence in the Q&A format during the dialogue sessions. The isolated individual unit of information is not disembodied knowledge. Rather, it models to students the potential to serve as evidence in relation to a claim that will fulfill a certain argumentative function (e.g., supports own side or weakens opposing side). In addition, students were also prompted to submit questions they would like answered.

By providing these pieces of evidence to students while they engaged in dialogues with the opposing side, students were thus prompted to act on the potential of using evidence for an argumentative goal. Moreover, when they construct an evidence-based claim, the strength and
validity of that claim is immediately under scrutiny from the opposing side. Counterarguments from the opposing side further prompt students to evaluate their initial evidence-based claim and make adjustments or provide a rebuttal. Thus, dialogues constitute an ideal context to provide students with rich information that has the potential to serve as evidence related to a topic.

The present study contributes to the learning to argue approach adopted in the current psycho-educational research on argumentation, given the main focus here is to develop students’ skills in constructing evidence-based claims that fulfill various argumentive functions, both in dialogic and individual settings. The findings that students were able to transfer their argument skills from one topic to another, and finally to a non-intervention topic, suggested that it is feasible to teach argument skills that are generalizable. We propose that an optimal approach to teach these skills is to engage students in extended argumentive dialogues with opposing-side peers.

Engaging students in the argument curriculum also helped them overcome, to a certain degree, the my-side or self-serving bias in evaluating belief-relevant evidence, a common bias frequently reported in numerous prior research (Lord et al., 1979; Kuhn, 1989, 1993). In the present study, 81% of students from the evidence reflection condition addressed evidence incongruent with their position during the last intervention topic, and 70% of them did so for a novel topic. Moreover, the majority of these students were able to coordinate a piece of evidence unfavorable to their position with another piece of evidence favorable to their position. These findings suggest that repeated interactions with a real-life other that argumentive dialogues provide, as is the case in the present study, make the opposing position and its accompanying evidence vivid enough to participants, so much so that they were able to temporarily decouple
from their position on the issue (Stanovich et al., 2013) and to consider and address evidence incongruent with their position.

In the present study, we advocate incorporating joint reflection into dialogic argumentation in order to achieve optimal results in developing students’ skills in making evidence-based claims in individual written arguments. Granted, measuring argumentive writing is a multifaceted construct and researchers devoted to writing research have been measuring it from different perspective (Graham & Perin, 2007). However, the focus of the present study is on developing and measuring students’ ability to coordinate claims with evidence, a skill we view as central to argument.

Limitations and future directions

Here we note four limitations of the present study. First, we did not measure student interest in the topics. The claim that students showed most interest for Topic 2 and struggled with Topic 3 was based on the researcher’s observation and anecdotal evidence. In future research, we could measure student interest for topics and see whether there is a correlation between interest and performance. Second, since the outcome measure in the present study is students’ individual essays at the end of the topic, we do not know how the two conditions differed in addressing evidence during the dialogues with the opposing side. It could be that the two conditions showed equivalent performance for all three topics but only the evidence reflection group transferred their skills from a dialogic setting to an individual setting. Alternatively, it could be that the evidence reflection group showed superior performance during dialogues, which would suggest that students who did not engage in evidence reflection did not even have the skills as solidly in place as the reflection condition before students conditions...
wrote individual essays. Thus, to decide between these alternatives, an important next step is to analyze the argumentive dialogues during each topic and examine dyads’ construction of evidence-based claims in these dialogues.

Third, it should be noted that the assumption that students who engaged in joint reflection developed better self-regulatory control over their evidence use in individual writing was not empirically tested in the present study, even though findings at the performance level did suggest that they were more meta-cognitively aware and meta-strategically in control of the demands of evidence use in constructing arguments. Thus, an important next step for future research is to test the effects of the manipulation on individuals’ self-regulation in individual context. We could ask students to “think aloud” while they read through the evidence list when writing an essay. We could then detect whether there is a greater presence of self-regulatory speech among students in the evidence reflection condition, and whether such speech is correlated with performance.

Lastly, since we did not audiotape dyads as they completed the evidence reflection sheets, we were not able to get a full picture of the extent to which they engaged with one another when carrying out reflective activities on evidence. What they actually wrote down on paper, an outcome measure used in the present study, could only represent a small, or even biased portion of what was under discussion. Thus, future research that incorporates peer collaboration could consider audiotaping and analyzing collaborative speech in order to better reveal the nature and developmental patterns of peer collaboration.
REFERENCES


Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and
Appendix A

QUESTIONS AND ANSWERS ABOUT JUVENILE & ADULT COURT

1. How does a juvenile court system differ from a regular one?
   The judges and staff in a juvenile system are specially trained to deal with young people who have committed crimes.

2. Are punishments for the same crime different in juvenile and adult courts?
   Yes, punishments tend to be less severe and sentences shorter in juvenile court. In 2006 in China, 70% of juveniles who committed serious crimes are subjected to sentences for less than 3 years in juvenile court.

3. At what age is the brain fully developed?
   The prefrontal cortex, which is responsible for abstract thinking and the ability to exercise good judgment, is not fully developed until one’s early- to mid-20s.

4. Would the government save money if they didn’t have to pay for a separate juvenile system?
   Yes, juvenile courts and prisons require more people to run and thus cost more. Adult courts cost less to operate.

5. Do juveniles ever commit violent crimes such as murder?
   Research shows an increase in violent crimes committed by juveniles in the past ten years in China. In particular, in 2012 in China, 5.1% of juvenile-committed crimes were murders.

6. Do teens sentenced to jail time in juvenile court get jail records?
   They do not if sentences are served in a juvenile detention center; their records are sealed on their release.

7. Do people have better self-control as they get older?
   Not necessarily. But there is evidence that adults think more carefully before they act, compared to teens.

8. Are teens or adults more likely to repeat their crimes?
   For teens convicted of a serious crime (a felony), the rate of recidivism (repeat crime) is 90% over 10 years. For crimes overall, it is about 50%.

9. What are public opinions on the juvenile court issue?
   A “get tough” policy has become more popular in recent years, with a law proposing that adolescents as young as 16 are tried in regular adult court.

10. Are teens at risk of being assaulted in adult prisons?
   Yes. Teens in adult jails are 50% more likely to be attacked by another inmate and twice as likely by prison staff, compared to adult prisoners.
11. Do adult jails provide job training?

Many do. In 2016, the majority of adult prisoners in China engage in labor reform which might train their job-related skills.
Appendix B  Argument Reflection Sheet

Team members __________________________
Date ____________________________

Let’s think...Starting with the other side’s argument

One of the other side’s MAIN ARGUMENTS was:

Our COUNTERARGUMENT against their argument was:

Give a specific example of an improved, more effective COUNTERARGUMENT.
Let's think...Starting with our argument

One of our MAIN ARGUMENTS was:

Their COUNTERARGUMENT against our argument was:

Our COMEBACK was:

How can this COMEBACK be improved?
Is there a more effective comeback?
Appendix C
Evidence distributed during each dialogue session for all topics

**Topic 1**

<table>
<thead>
<tr>
<th>Chat 1</th>
<th>Only Schoolwork</th>
<th>Part-time job</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence #1</td>
<td>M+</td>
<td>Evidence #2</td>
</tr>
<tr>
<td>Q: Does high school require a lot of time devoted to schoolwork?</td>
<td>M+</td>
<td>Q: Do students need work experience before they can decide what career they want to prepare for?</td>
</tr>
<tr>
<td>A: Most teachers who have taught high school in China say that high school curriculum can be very demanding and students often have a lot of homework after school.</td>
<td></td>
<td>A: Studies have shown that students who have work experience develop clearer ideas of what they want to do in the future.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chat 2</th>
<th>Evidence #3</th>
<th>Evidence #4</th>
</tr>
</thead>
<tbody>
<tr>
<td>O-</td>
<td>Q: Can teenagers learn bad habits in work settings?</td>
<td>O-</td>
</tr>
<tr>
<td>A: Yes, some parents of teenagers who have worked part time noticed that their kids started to imitate bad habits of older people they work with.</td>
<td></td>
<td>A: No, a recent opinion poll shows that some high school students want to prepare for jobs rather than more schooling.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chat 3</th>
<th>Evidence #2</th>
<th>Evidence #1</th>
</tr>
</thead>
<tbody>
<tr>
<td>O+</td>
<td>Q: Do students need work experience before they can decide what career they want to prepare for?</td>
<td>O+</td>
</tr>
<tr>
<td>A: Studies have shown that students who have work experience develop clearer ideas of what they want to do in the future.</td>
<td></td>
<td>A: Most teachers who have taught high school in China say that high school curriculum can be very demanding and students often have a lot of homework after school.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chat 4</th>
<th>Evidence #4</th>
<th>Evidence #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-</td>
<td>Q: Do all high school students want to pursue an academic track leading to college?</td>
<td>M-</td>
</tr>
<tr>
<td>A: No, a recent opinion poll shows that some high school students want to prepare for jobs rather than more schooling.</td>
<td></td>
<td>A: Yes, some parents of teenagers who have worked part time noticed that their kids started to imitate bad habits of older people they work with.</td>
</tr>
</tbody>
</table>

| Chat 5     | Evidence #5     | Evidence #6   |
**M+**
Q: Do high school students who spend more time on schoolwork have a better chance to get into a good college?
A: Yes, given the competitive nature of the college entrance examination in China, historical data show that students who study more are more likely to score higher and have a better chance to get into a good college.

**M+**
Q: Do high school students develop new useful skills when doing a part-time job?
A: Yes, studies have shown that students who work part-time might be more likely to develop important skills than other students, such as being more accountable and enhanced ability to deal with an emergency situation.

**Topic 2**

<table>
<thead>
<tr>
<th>Position</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chat 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence #1</td>
<td>M+</td>
<td></td>
</tr>
<tr>
<td>Q: Why have animals been used in medical research?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: Animal organs often resemble human organs, so medicines may work in similar ways.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence #2</td>
<td>M+</td>
<td></td>
</tr>
<tr>
<td>Q: Can models made out of human tissue be used to test new drugs?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: Recently developed models made from human cells were designed to mimic more closely how the human body works and can be used to test how our bodies will react to drugs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chat 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence #3</td>
<td>O-</td>
<td></td>
</tr>
<tr>
<td>Q: Are there any types of research that could be performed with animals but not humans?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: Many studies of living bodies are so complicated and uncertain that they could only be carried out with animals. For example, studies in gene engineering test how to modify the organs of animals so they can be transplanted to humans.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence #4</td>
<td>O-</td>
<td></td>
</tr>
<tr>
<td>Q: Do most of the drugs that pass animal tests succeed in humans?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: The Food and Drug Administration reports that 92 out of every 100 drugs that pass animal tests fail in humans.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chat 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence #2</td>
<td>O+</td>
<td></td>
</tr>
<tr>
<td>Q: Can models made out of human tissue be used to test new drugs?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: Recently developed models made from human cells were designed to mimic more closely how the human body works and can be used to test how our bodies will react to drugs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence #1</td>
<td>O+</td>
<td></td>
</tr>
<tr>
<td>Q: Why have animals been used in medical research?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: Animal organs often resemble human organs, so medicines may work in similar ways.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chat 4  
Evidence #4  
M-  
Q: Do most of the drugs that pass animal tests succeed in humans?  
A: The Food and Drug Administration reports that 92 out of every 100 drugs that pass animal tests fail in humans.

Evidence #3  
M-  
Q: Are there any types of research that could be performed with animals but not humans?  
A: Many studies of living bodies are so complicated and uncertain that they could only be carried out with animals. For example, studies in gene engineering test how to modify the organs of animals so they can be transplanted to humans.

Chat 5  
Evidence #5  
M+  
Q: Has animal testing led to cures for any human diseases?  
A: Animal testing has led to treatments and cures for many human diseases. For example, research with dogs led to treatments for diabetes, research with armadillos led to leprosy vaccines, and research with monkeys have led to treatments for hepatitis, polio, and AIDS.

Evidence #6  
M+  
Q: Can synthetic versions of human organs be used in research?  
A: Studies involving the effect of sunscreen on a material like human skin gave quick results, compared to the length of time required for animal testing.

## Topic 3

<table>
<thead>
<tr>
<th>Position</th>
<th>Sell</th>
<th>Donation</th>
</tr>
</thead>
</table>
| Chat 1   | Evidence #1  
M+  
Q: Do people die because they can't get a new kidney in time?  
A: Yes, statistics show that in China in 2012, for every 150 people waiting for a donated kidney, 149 die while waiting.  | Evidence #2  
M+  
Q: Have many people agree to donate a kidney?  
A: Currently, a certain percentage of Chinese choose to be organ donors. However, France has increased their donors to 99% by assuming that everyone wants to donate their organs unless they notify in writing that they don’t want to (this is called “opting out”). |
| Chat 2   | Evidence #3  
O-  
Q: Do enough people volunteer to donate their kidneys for there to be enough kidneys to go around to those who need them?  | Evidence #4  
O-  
Q: How much do kidneys sell for?  
A: Economists estimate that kidneys can cost anywhere from 30,000 to 40,000 RMB |
| Chat 3 | Evidence #2  
| O+  
| Q: Have many people agree to donate a kidney?  
| A: Currently, a certain percentage of Chinese choose to be organ donors. However, France has increased their donors to 99% by assuming that everyone wants to donate their organs unless they notify in writing that they don’t want to (this is called “opting out”). |
| Evidence #1  
| O+  
| Q: Do people die because they can’t get a new kidney in time?  
| A: Yes, statistics show that in China in 2012, for every 150 people waiting for a donated kidney, 149 die while waiting. |

| Chat 4 | Evidence #4  
| M-  
| Q: How much do kidneys sell for?  
| A: Economists estimate that kidneys can cost anywhere from 30,000 to 40,000 RMB if selling is allowed. This is more money than the average annual income of a family in China. |
| Evidence #3  
| M-  
| Q: Do enough people volunteer to donate their kidneys for there to be enough kidneys to go around to those who need them?  
| A: No. Currently in China there are about a million people on waiting list to receive kidney transplants. However, only about 10,000 kidney transplant surgeries are carried out each year. |

| Chat 5 | Evidence #5  
| M+  
| Q: Do people who sell their kidneys need the money they receive?  
| A: Yes. Almost always they are very poor and have few ways to earn money. For instance, some kidney sellers in Iran need the money to pay off debts. |
| Evidence #6  
| M+  
| Q: How many people volunteer to be organ donors in China?  
| A: In 2016, over 100,000 people registered to be organ donors, a threefold increase over 2015. Since the Department of Health in China opened up online self-registration in 2010, the number of registered donors has dramatically increased each year. |
Appendix D
Evidence lists distributed to students during essay writing for all topics

Topic 1
1. Q: Does high school require a lot of time devoted to schoolwork?
   A: Most teachers who have taught high school in China say that high school curriculum can be very demanding and students often have a lot of homework after school.

2. Q: Do students need work experience before they can decide what career they want to prepare for?
   A: Studies have shown that students who have work experience develop better ideas of what they want to do in the future.

3. Q: Can teenagers learn bad habits in work settings?
   A: Yes, some parents of teenagers who have worked part-time noticed that their kids started to imitate bad habits of older people they work with.

4. Q: Do all high school students want to pursue an academic track leading to college?
   A: No, a recent opinion poll shows that some Chinese high school students want to prepare for jobs rather than more schooling.

5. Q: How long are the summer and winter breaks for high school students in China?
   A: While different schools have different break schedules, most high schools schedule extra school time during breaks, so the length of their summer and winter breaks might get cut in half.

6. Q: What do parents from Western countries (such as the US) think about part-time jobs for their children?
   A: While different Western parents hold different opinions, statistics have shown that the percentage of Western high school students who hold a part-time job is higher than that of Chinese high school students.

7. Q: What kind of part-time jobs can high school students take in China?
   A: There is a range of jobs available to high school students. Some work in the service industry like a cashier. Some provide tutoring to younger students. In recent years, companies such as Google offer paid internship opportunities for high school students.

8. Q: How much can high school students earn when doing a part-time job?
   A: High school students and adults who work in the same sector in the service industry earn about the same. In 2017, the average monthly wage of a full-time employee at Walmart is around 3000 RMB and the hourly wage is about 12 RMB.

9. Q: Do high school students develop new useful skills when doing a part-time job?
   A: Yes, studies have shown that students who work part-time might be more likely to develop important skills than other students, such as being more accountable and enhanced ability to deal with an emergency situation.
10. Q: Do high school students who spend more time on schoolwork have a better chance to get into a good college?
A: Yes, given the competitive nature of the college entrance examination in China, historical data show that students who study more are more likely to score higher and have a better chance to get into a good college.

11. Q: Can high school students who do a part-time job receive mistreatment?
A: Yes that is possible. According to a recent research project in Japan, 70% of Japanese high school part-time workers have received mistreatment at least once, such as unpaid overtime work.

12. Q: What is the yearly tuition and living expense for a high school student in China?
A: With the increase in price over the past few years, the cost for attending high school has dramatically increased. For students enrolled in a private high school, their parents need to pay as much as 50,000 RMB per year for tuition and living expenses combined.

Topic 2

1. Q: How many animals are involved in medical research each year in the USA?
A: The U.S. Department of Agriculture reports that 1.2 million animals were used in research 2005. This does not include rats and mice, which make up about 90% of animals used in research.

2. Q: Can researchers use as many animals as they want in their research?
A: Regulations exist that require that scientists use as few animals as possible to conduct their research.

3. Q: Can bodies of humans who recently died be used for research?
A: Examining human bodies soon after death can help to better understand causes and effects of diseases and medicines.

4. Q: Why have animals been used in medical research?
A: Animal organs often resemble human organs, so medicines may work in similar ways

5. Q: Do most of the drugs that pass animal tests succeed in humans?
A: The Food and Drug Administration in America reports that 92 out of every 100 drugs that pass animal tests fail in humans.

6. Q: Are there any types of research that could be performed with animals but not humans?
A: Many studies of living bodies are so complicated and uncertain that they could only be carried out with animals. For example, studies in gene engineering test how to modify the organs of animals so they can be transplanted to humans.

7. Q: Can synthetic versions of human organs be used in research?
A: Studies involving the effect of sunscreen on a material like human skin gave quick results, compared to the length of time required for animal testing.
8. Q: Has animal testing led to cures for any human diseases?
A: Animal testing has led to treatments and cures for many human diseases. For example, research with dogs led to treatments for diabetes, and research with monkeys have led to treatments for hepatitis, polio, and AIDS.

9. Q: How are animals treated in research laboratories?
A: There are laws in place to help ensure that distress and pain in animals is kept to a minimum, but the daily treatment of animals is not known because the testing places cannot be monitored at all times and records are not shared.

10. Q: How similar are humans and animals in terms of diseases they get?
A: Many of the diseases that humans get—such as cancer, malaria, asthma, arthritis, and heart failure—are also found in animals.

11. Q: Can statistics be used to analyze how people react to different life events?
A: Statisticians have helped link cigarette smoke to lung cancer and diet to heart disease by studying large numbers of people over periods of time.

12. Q: Can medical testing of animals be of any benefit to animals?
A: Many of the medications that are given to sick animals (such as pets and zoo animals) were discovered as a result of medical research with humans that involved those animals.

Topic 3

1. Q: Do people die because they can't get a new kidney in time?
A: Yes, statistics show that in China in 2012, for every 150 people waiting for a donated kidney, 149 die while waiting.

2. Q: Have many people agree to donate a kidney?
A: Currently, a certain percentage of Chinese choose to be organ donors. However, France has increased their donors to 99% by assuming that everyone wants to donate their organs unless they notify in writing that they don’t want to (this is called “opting out”).

3. Q: Do enough people volunteer to donate their kidneys for there to be enough kidneys to go around to those who need them?
A: No. Currently in China there are about a million people on waiting list to receive kidney transplants. However, only about 10,000 kidney transplant surgeries are carried out each year.

4. Q: How much do kidneys sell for?
A: Economists estimate that kidneys can cost anywhere from 30,000 to 40,000 RMB if selling is allowed. This is more money than the average annual income of a family in China.

5. Q: Will taking out one kidney negatively affects the health?
A: No, the negative effect of having one kidney removed is very limited. Some people are even born with just one kidney and lead a normal life. As long as the other kidney is healthy, you are expected to have a normal life expectancy.

6. Q: How many people volunteer to be organ donors in China?
A: In 2016, over 100,000 people registered to be organ donors, a threefold increase over 2015. Since the Department of Health in China opened up online self-registration in 2010, the number of registered donors has dramatically increased each year.

7. Q: How well does the legal kidney market work in Iran?
A: Since Iran’s legalization of the kidney market in the mid-90s, the need of kidney transplant has been met. The sellers and recipients would match their blood type and run other tests in large hospitals before the transplant surgery is carried out.

8. Q: Do people who sell their kidneys need the money they receive?
A: Yes. Almost always they are very poor and have few ways to earn money. For instance, some kidney sellers in Iran need the money to pay off debts.

9. Q: What is the percentage of kidneys coming from the black market?
A: According to the statistics in America, in 2010, 1/5 of the kidneys come from the black market.

10. Q: Is it easy to make known your wish to donate your organs when you die?
A: Very easy. In America, many states encourage donations by allowing the consent to be noted on a person’s driver’s license. In China, you simply need to go onto a website and register as an organ donor.

11. Q: For patients with kidney failure, are there other treatment plans besides kidney transplant?
A: Patients with kidney failure could stay on dialysis. However, dialysis is costly. A patient needs 2-3 procedures each week with each procedure costs around 800 RMB without insurance.

12. Q: Can a kidney get transplanted from the body of someone who has died?
A: Yes, if it is done quickly after death and the donor’s family agrees.