Alberta Journal of Educational Research, Vol. 68.4, Winter 2022, 463-479

Elementary Students' Argument Evaluation in a Science Classroom

Qingna Jin

University of Alberta

Students' ability to evaluate arguments is significant in democratic societies. Therefore, researchers argue that it is important to understand and facilitate students' argument evaluation skills. Most research on students' argument evaluation focuses on final products and outcomes instead of students' decision-making processes. Thus, how students evaluate arguments constructed by others has not been fully clear to researchers and educators. This qualitative case study aimed to understand the process of students' argument evaluation by exploring the affordance of a new data collection method. In addition to the written task, which is commonly used in the current research on argument evaluation, this study also employed stimulated recall interviews (SRIs) to access students' inner awareness and thinking processes while they were engaged in the argument evaluation task. Data collected with SRIs were analyzed qualitatively, together with students' written responses. Findings from this study reveal that students' argument evaluation is a complex cognitive endeavor, and the actual process of argument evaluation is more sophisticated than what is demonstrated in students' written responses. Based on these findings, this study suggests that only examining students' written products might not be sufficient to achieve a comprehensive understanding of their argument evaluation skills. Considerations of using SRIs are also discussed.

Dans les sociétés démocratiques, la capacité des élèves à évaluer des arguments est importante. Les chercheurs soutiennent qu'il est donc important de comprendre et de faciliter les compétences des élèves en matière d'évaluation des arguments. La plupart des recherches sur l'évaluation des arguments par les élèves portent sur les produits et les résultats finaux plutôt que sur les processus décisionnels des élèves. Les chercheurs et les éducateurs n'ont donc pas une idée claire de la manière dont les élèves évaluent les arguments formulés par d'autres. Cette étude de cas qualitative visait à comprendre le processus d'évaluation des arguments par les élèves en explorant les possibilités d'une nouvelle méthode de collecte de données. En plus de la tâche écrite, qui est couramment utilisée dans la recherche actuelle sur l'évaluation des arguments, cette étude a également utilisé des entretiens de rappel stimulé pour accéder à la conscience intérieure et aux processus de pensée des élèves pendant qu'ils étaient engagés dans la tâche d'évaluation des arguments. Les données recueillies au moyen des entretiens de rappel stimulé ont été analysées de manière qualitative, ainsi que les réponses écrites des élèves. Les résultats de cette étude révèlent que l'évaluation des arguments par les élèves est un effort cognitif complexe, et que le processus réel d'évaluation des arguments est plus sophistiqué que ce qui est démontré dans les réponses écrites des élèves. Sur la base de ces résultats, cette étude suggère que le seul examen des textes écrits des élèves pourrait ne pas être suffisant pour obtenir une compréhension complète de leurs compétences en matière d'évaluation des arguments. Les considérations relatives à l'utilisation des entretiens de rappel stimulé sont également discutées.

Argumentation is a fundamental skill for the 21st century citizens of democratic societies (Chen, 2020; Cavagnetto, 2010; Duschl & Osborne, 2002; Kuhn, 2019). Evaluating an argument is even more important than constructing one, because people are called on more frequently to evaluate arguments in their everyday life. To make an informed decision, people usually need to compare various options by examining the available information and evidence, instead of making judgments based on personal intuition or gut feelings. Therefore, one of the goals of education is to teach students to critically evaluate arguments constructed by others to enable them to make informed personal and social decisions (Breakstone et al., 2018; Kuhn, 2019; Osborne et al., 2004). However, students' argument evaluation process has not been clear to educators yet. This is because many studies on argument evaluation focus on final products and outcomes, such as whether students can identify valid or trustworthy arguments (Chen et al., 2016). Limited research has explored the process of how students make their judgments. Moreover, the majority of the research on argument evaluation has focused on secondary and higher levels of education, while relatively fewer studies have explored students' argument evaluation at the elementary level. This study focused on elementary students' argument evaluation process in a science classroom.

Background

Scientific Argument and Students' Argument Evaluation

Scientific discourse is characterized by rational dispute and the exchange of arguments. Therefore, the ability to comprehend and assess the validity of scientific arguments is an essential facet of scientific literacy (Eugenio-Gozalbo et al., 2022; Lytzerinou & Iordanou, 2020; Münchow et al., 2019). Scientific texts usually contain informal arguments that are different from formal deductive arguments, in which a conclusion follows with logical necessity from a set of promises. Toulmin's (1958) model of an argument has been widely adopted in research on students' scientific argumentation. This study adopted a Toulmin-inspired framework of informal argument. The framework is similar to ones adopted by a number of other researchers in science education and has proven helpful to interpret student argumentation (see, e.g., Chen, 2020; Guilfoyle & Erduran, 2021; Hosbein et al., 2021; McNeill et al., 2016; Osborne et al., 2004; Sampson & Clark, 2009). Within this framework, the structure of a scientific argument consists of a *claim* supported by *evidence* and *reasoning*. A claim is a tentative statement that provides an answer to a certain question; evidence, in its broadest sense, includes anything such as measurement or observation that is used to support the validity or legitimacy of the claim; and reasoning is the justification explaining how the evidence supports the claim (Sampson & Clark, 2009). The reasoning component of the framework indicates that a strong argument needs to include not only evidence, but also a rationale that shows why the evidence supports the claim and why the evidence provided should count as evidence (Chen, 2020; Sampson et al., 2011). For distinguishing strong from weak arguments, students need to consider not only the accuracy of the provided information, but also the completeness and the internal consistency of the arguments (Shaw, 1996). Thus, exploring how students evaluate arguments has the potential to reveal the epistemological criteria that they believe govern reasoning, such as what counts as valid evidence (Cheng & Yang, 2022; Kuhn et al., 2013; Ryu & Sandoval, 2012). Understanding how students perceive and appreciate epistemic norms of scientific reasoning is important for the development of their argumentation skills and science learning (Sampson & Clark, 2008, 2009).

Research on students' argument evaluation tends to ask students to assess arguments with written tasks (see, e.g., Cheng & Yang, 2022; Kuhn et al., 2013; Münchow et al., 2019; Ryu & Sandoval, 2012). Usually, students are provided with different arguments and asked to write down which one(s) they think is/are better or stronger and explain reasons for their decisions. Then, students' written responses are collected and analyzed, either quantitatively or qualitatively, to assess their abilities of argument evaluation, to reveal their epistemological beliefs, and/or to identify if there is any progress after certain instructional interventions. These analyses of students' written responses have reported varied results. For example, some studies have reported that students usually focus on the correctness of a claim, rather than looking at the relationships among question, claim, and evidence (see, e.g., Choi et al., 2010; Takao & Kelly, 2003). Kuhn et al. (2013) further argued that students often tend to do this because they are inclined "to simply evaluate the content of a statement" (p. 474). However, elementary students in Ryu and Sandoval (2012), even at the pretest before the instructional intervention, appropriated epistemological criteria for argument evaluation centered around evaluations of evidence and the fit between evidence and claims.

Many factors contribute to the varied findings in the literature about students' argument evaluation (Iordanou et al., 2016). For example, each study has its unique research context, and student participants in these studies are different in terms of, for example, grade levels and previous learning experiences. Nevertheless, these coexisting varied results have raised a debate on whether young students are or are not capable of critical argument evaluation, and thus suggest that more research is needed to further enrich our understandings about students' argument evaluation. Whether young learners are capable of participating in and benefiting from critical argument evaluation raises further significant questions in terms of what kinds of instructional support are appropriate and helpful to facilitate students' argument evaluation. The prerequisite for designing or implementing any instructional intervention is that researchers and educators comprehensively understand students' argument evaluation as a complicated cognitive, epistemic, and social process (Kim & Roth, 2014; Settlage & Southerland, 2019). Therefore, instead of relying on one single technique (e.g., collecting and analyzing students' written responses to the argument evaluation tasks), researchers should consider employing multiple methods, which could supplement and triangulate each other, to form a comprehensive view on students' argument evaluation. These attempts have the potential to enrich our theoretical understandings about students' argument evaluation and inform the design and implementation of appropriate instructional support to facilitate students' critical argument evaluation, as well as to explore more methodological possibilities to investigate student argumentation.

To contribute to these issues, this study adopted stimulated recall interview (SRI) together with written tasks to explore how elementary students make their decisions during the process of comparing and evaluating various arguments. With the SRI, the study particularly focused on students' inner awareness and thinking process of making their judgments, which have usually been neglected in the existing literature.

Challenges to Accessing Students' Inner Awareness and Thinking Processes

The challenge of approaching people's inner awareness and thinking processes has been well recognized and documented (Thomas & McRobbie, 2001). It is difficult, if even possible, to observe feelings, thoughts, or how people interpret the world around them. Thus, researchers use self-report methods, such as interviewing people, to find out from them those things that cannot

be directly observed (Merriam, 2016; Patton, 2015). To collect relevant data on students' thinking, researchers have tried many self-report methods, yet each of these methods has inherent pros and cons (Anderson et al., 2009). For example, with questionnaires or surveys, some studies invite students to predict or recall the process of completing certain cognitive tasks (Veenman, 2005). Questionnaires or surveys can be easy to administer to large groups of students, but it has been noticed that scores on these questionnaires have a low correspondence with behavioral or cognitive measures during task performance (Veenman, 2005; Veenman et al., 2006). Think-aloud protocol is another method that is widely used to approach students' thinking (Anderson et al., 2009). With the think-aloud strategy, researchers gather students' verbal reports of their thought processes while they are completing a task such as solving a problem (Branch, 2006). Think-aloud conversations are usually conducted individually with student participants and have the potential to provide researchers with rich qualitative data (Branch, 2006). However, this kind of conversation may prevent students from learning the presented materials and may interrupt their task performance since students are asked to express their inner thinking process verbally while they are performing the task (Akturk & Sahin, 2011; Malva et al., 2021). Likewise, self-report interview as a method to approach students' inner thinking process also has its own strengths and limitations. The most important strength of interviews is that they can provide rich and highly qualitative data on how students think and how they interpret the world from their own perspectives (Anderson et al., 2009; Deppermann, 2013). Moreover, interviews offer researchers opportunities to ask further questions if an answer is incomplete (Veenman, 2005). However, it is noteworthy that, when students are interviewed to report their thinking processes, it usually requires them to reconstruct or reflect on these thinking processes. Such reconstructions or reflections may distort the thinking processes reported (Veenman, 2005). Also, students might know more than they tell, and they sometimes tell more than they know.

To circumvent the aforementioned potential memory failures or distortions, scholars (e.g., Malva et al., 2021; Rowe, 2009; Veenman, 2005) have suggested using stimulated recall interviews (SRI). During SRI, student participants are invited to review their products and/or performances of a specific task to reproduce their thought process and then share with the researchers their thinking at that time (Thomas & McRobbie, 2001; Veenman, 2005). In this way, SRI provides researchers a rich and highly qualitative means to elucidate students' thinking processes and awareness within the context in which they are situated (Anderson et al., 2009). This study adopted SRI as the main method, together with a written task, to collect relevant data on students' argument evaluation process.

This Study

Qualitative Case Study

This paper was a part of a larger project on student argumentation which investigated how students learn science through participating in argumentative practice, including argument construction, argument evaluation, and dialogical argumentation (i.e., the dialogical form of argumentation). This paper focused on students' argument evaluation. As a part of a qualitative cast study, this paper was descriptive in nature (Merriam, 2016). With its descriptive emphasis, the paper presented a detailed account of the phenomenon under study, that is, elementary students' argument evaluation in their science classroom. This account is useful because educators have not fully understood how students make judgments in terms of comparing and

evaluating various arguments (Merriam, 2016). Moreover, this paper described how SRI helped the researcher gather relevant data to understand students' argument evaluation but influenced students' thinking and task performance. These results may be helpful for researchers who are considering employing this method (i.e., SRI) in future studies.

A Grade 5/6 science class in a public school in Canada participated in this project during the four-month long research period which covered two learning units. Nineteen student participants included four girls and fifteen boys, twelve of whom were in Grade 6 and seven in Grade 5. Multiple methods for data collection were employed in the project, including classroom observation, interviews, and collecting students' writings. The main data sources on students' argument evaluation which is the focus of this paper included students' written responses to the argument evaluation task and the stimulated recall interviews with them, which are described in detail in the following section.

The Argument Evaluation Task

An argument evaluation task was designed for the research purpose of this study. All the nineteen student participants were invited to complete this task. This task happened in the middle of the research period and between the two learning units. When they wrote this argument evaluation task, they had completed the unit of Sky Science and learned the properties of air (e.g., air has mass and weight). Within the unit of Sky Science, they had experienced argument construction and dialogical argumentation. With the teacher's scaffolds, they had developed some basic understandings about norms of argumentation, such as evidence is needed and helpful to make a claim (see Jin & Kim, 2021). Within this argument evaluation task, students were asked to evaluate five different arguments. All of these arguments were constructed to answer and explain the questions "Which soccer ball is heavier, the inflated one or the flat one? And why?"

This written task had two parts. In the first part, students were asked to answer the question "Which argument do you think is the *most* convincing? And why?" That is, they were to choose the most convincing argument from the following four:

> Argument A: The inflated one is heavier because it has both leather and a lot of air.
> Argument B: The flat one is heavier. I play soccer and I can't kick the flat ball as far as the inflated one, so the flat one must be heavier.
> Argument C: The flat soccer ball is mostly leather and has very little air in it, and leather is heavier than air, so the flat soccer is heavier.
> Argument D: Both soccer balls weigh the same because air does not weigh anything.

These four arguments made different claims (i.e., different viewpoints in terms of whether the inflated or the flat soccer ball was heavier) and corresponding evidence. In the second part, students were asked to decide which argument was more convincing between the following two that made the same claim (i.e., the inflated soccer ball is heavier than the flat one) yet offered different justifications and evidence.

> Argument A: The inflated one is heavier because it has both leather and a lot of air.
> Argument E: The inflated soccer ball is heavier because it has the same amount of leather as the flat one, and at the same time, more air in it. Air has weight, so the inflated one is heavier than the flat one.

In both parts, students were asked to write down their choices and the reasons for their decisions. Students completed this argument evaluation individually.

Stimulated Recall Interview (SRI)

This study adopted SRI to probe students' thinking processes of making their decisions. During SRIs, students reviewed their written responses and sometimes viewed video clips of their argument evaluation to recall their thought processes (Thomas & McRobbie, 2001). Then, with the researcher's prompts, students verbally expressed how they were thinking at that time. Sometimes they also clarified their answers and explained rationales for their choices during the interviews. Eight students, including seven boys and one girl, volunteered to participate in the SRIs. The selection of the interviewees was mainly based on their willingness to share their ideas and thoughts with the researcher. The small sample has the potential to represent the whole class in terms of their argument evaluation. When the interviewees' written responses were examined, their answers covered all the choices that the entire class had made. For instance, in the first part of the task, different numbers of students in the class had chosen A, B, and D as their answers (the following Results section includes detailed descriptions of these three groups), and these eight interviewees included students from all these three groups. Nevertheless, it is suggested to be cautious whenever generalization is made from the small sample to a broader context (Guba & Lincoln, 1994).

Students were interviewed individually. Each interview was around 40 to 50 minutes long and was video recorded. These eight SRIs took place over two days right after the students completed the argument evaluation task to minimize memory failure. Each interview started with casual talks about what interested the student (e.g., "Was there anything interesting that happened when you think back about doing this task or learning the Sky Science unit?") and then gradually moved to questions related to their thought processes, such as "I noticed that you changed your answer. Why did you change it and what were you thinking at that time?" and "Would you like to share with me how you thought to make your choice?"

Bernard (1994) suggests qualitative researchers "being [at] a site over a period of time" to familiarize themselves with the community under investigation and thereby "facilitate involvement in sensitive activities to which [they] generally would not be invited" (p. 142). Students' inner thinking and awareness are sensitive in nature; therefore, this study adopted Bernard's suggestion. Specifically, before the students wrote the argument evaluation task and participated in the subsequent SRIs, I had been in the classroom for several weeks. I observed the science teaching and learning taking place in the classroom, such as interactions between students and their teacher and peers. These observations familiarized me with the research context and thus helped me better interpret the student interviews. Moreover, the time that I spent in the classroom before the actual data collection also familiarized the students with my presence and was helpful for establishing and maintaining rapport between the researcher and student participants (Bogdan & Biklen, 2007). In this way, I managed to have students feel safe and be willing to share their thinking processes with me during the SRIs.

Data Analysis

Data analysis was a multi-step and iterative process. A colleague in science education was invited to participate in the data analysis for the purpose of peer checking. First, all of the interview

transcripts, students' written responses, and the researcher's field notes were viewed for open coding to generate initial themes. During the subsequent data-sharing and debriefing sessions with the colleague (i.e., axial coding), the initial themes together with their supporting data were discussed. There was some degree of disagreement, which mainly clustered around the influence of SRI as a method for data collection on students' thought processes. These differences and disagreement were solved by intensive discussion and revision through the interactive video analysis approach (Jordan & Henderson, 1995). With interactive video analysis, peers view and interpret video clips together by critically and creatively examining each other's interpretations and themes in order to reach consensus on the data analysis (Kim, 2016). With axial coding, we disassembled and reassembled the data to construct an overall understanding of themes. For selective coding, we selected a few episodes which distinctively demonstrate the themes that we agreed upon collectively. During this coding, we looked into the actual data of students' verbal and written explanations to develop the depth of themes. Findings that emerged from these processes of analysis are presented and discussed next.

Results and Discussion

During the stimulated recall interviews (SRIs), students described their thought processes of evaluating different arguments. These descriptions, together with their written responses in the argument evaluation task, provided a comprehensive understanding of students' argument evaluation process. Some inconsistencies emerged when students' written answers were cross-checked with the SRIs. It was noticed that students' actual thinking processes of evaluating arguments were more complex than what was represented in their written answers. Analyses of all these data, especially the SRIs, revealed that students' argument evaluation is a complex cognitive endeavor involving their epistemic understandings. SRI as a method to access students' inner thinking processes provided informative data on students' argument evaluation. Yet, it was noticed that SRI as a data collection method was not neutral; instead, it influenced students' performances and thought processes by engaging students in active reflection on their own thinking. In what follows, I describe these notions with illustrating examples.

Evaluating Arguments with Different Claims

Results from Analysis of Students' Written Responses

All 19 students wrote this argument evaluation task, and their written responses were collected. Two students did not write any further explanation of their choices; thus, their responses were excluded from the later analysis. In the first part of the task, students were asked to choose the most convincing argument from four that had different claims (see above in the previous section). Students' written responses demonstrated that all of them (n = 17) agreed and shared the same viewpoint with the argument they chose. In other words, even if they made different choices regarding the most convincing argument (12, 3, 0, and 2 students chose argument A, B, C, and D, respectively, as the most convincing one), they tended to agree with the claims in the arguments they chose. For example, Oliver (Oliver is a pseudonym, as are all the student names in this paper) chose Argument B as the most convincing one and agreed that "the flat [soccer ball] is heavier," which is the claim in Argument B. Isaac thought Argument A was the most convincing, and he also agreed with Argument A's claim that "the inflated [soccer ball] is heavier [than the flat one]."

Students were asked to write down the reasons for their decisions. Analyzing their written reasons revealed some patterns. Some students (10 out of 17) tended to repeat the claim that they agreed with when they explained reasons for their choices. For example, Ming thought Argument D (i.e., both soccer balls weigh the same because air does not weigh anything) was the most convincing because, as he wrote in his handout, "the inflated soccer ball and the flat one have the same weight, no difference." Oliver chose Argument B, which argued that the flat soccer ball was heavier, as the most convincing one; in his written response explaining the reasons, he wrote, "If I have an inflated soccer ball, it is lighter than the deflated one." Other students (7 out of 17) tended to supplement the evidence to further justify the claim they agreed with. For example, Isaac thought Argument A (i.e., the inflated soccer ball is heavier because it has both leather and a lot of air) was the most convincing one. When he was asked to explain the reasons for his decision, he wrote, "because I agreed with this, because they both have the same amount of leather, and the inflated one also has air, if air has weight, so the inflated soccer ball is heavier." Similarly, Tian explained that he thought Argument D was the most convincing because "these two balls weigh the same, air does not weigh anything, and it is like folding clothes, it will not change the weight of the jacket." They supplemented the evidence and further justified the claim in the chosen argument.

From the analysis, it seems that students thought the argument with the claim that they agreed with was the most convincing one. In other words, it seems as if students only examined whether the claims in the arguments were correct or not as they made their decisions about whether the argument was convincing or strong. This is similar to the results reported in previous studies that students usually focus on the correctness of a claim, rather than looking at the relationships among question, claim, and evidence (Choi et al., 2010; Takao & Kelly, 2003). Scholars such as Kuhn et al. (2013) have argued that it indicates that students have poor argument evaluation competence if they evaluate claims alone without determining whether available evidence is valid, relevant, sufficient, and convincing enough to support the claim. According to Kuhn et al. (2013), students often tend to do this because they are inclined "to simply evaluate the content of a statement" (p. 474). In their study, Kuhn and colleagues categorized students' performance in argument evaluation as "evaluation of claim alone" if students only examined whether the content of the claims was correct or whether they agreed with the statements or not.

Results from Analysis of SRIS With Students

In addition to students' written responses, SRIs with students were also analyzed. Analyzing these interviews revealed that, even if students eventually chose the argument with the correct claim (they thought) as the most convincing one, their thinking processes were diverse and complicated. Students' epistemic understandings about such things as what counts as evidence and what makes a claim were usually involved in the process of evaluating these provided arguments. The following two examples of Eli and Oliver can illustrate this. In these examples, both boys chose the argument with the "correct" claim as the most convincing one. Regarding the question "Which soccer ball is heavier, the inflated one or the flat one?" they had different answers. Therefore, their choices of the most convincing argument were different.

Eli thought argument A was the most convincing and explained in his handout that "the inflated soccer ball is heavier … because it has air and leather, and the flat [soccer ball] has no air, so it is only leather which makes it lighter." During the SRI and with the researcher's questions, Eli elaborated his thought processes as follows.

1-1 Eli: A is most convincing, because A is correct. You know, air has weight, so the inflated one is heavier.
1-2 Interviewer: How did you know air has weight?
1-3 Eli: I learned from science class, you know, air has weight. I know it. That is correct and convincing.
1-4 Interviewer: So, you think the argument with the correct answer is the most convincing, right?
1-5 Eli: … Yeah, I thought in that way. The answer should be correct. If it is not correct, it cannot be convincing … To be convincing, to be [a] good and strong argument, you need to be correct … so I chose A [as] the most convincing one.
…
1-6 Eli: But A did not really explain well, like, "a lot of"—what does it mean? It should clearly say that it is because air has weight, just like E, then, it will be much better … I considered these, I mean I knew A didn't really explain clearly, but others were wrong.

While Eli was making his judgment, he recalled the knowledge he had learned from previous science classes (i.e., "air has weight") (Turn 1-3), so he confidently knew "the inflated one is heavier" (Turns 1-1, 1-3); that is, he knew Argument A had the correct claim and agreed with it. As he continued sharing his thought process, he demonstrated his epistemic understandings about what makes a good argument and what convincing means. He believed that "if [the claim] is not correct, it [the argument] can't be convincing" and to be strong and convincing, the argument needs to have a correct claim (Turn 1-5). Based on what Eli shared thus far, it seemed his argument evaluation was "evaluation of claim alone," which is seen as an indicator of "poor [argument] evaluation" (Kuhn et al., 2013, p. 475). However, Eli's argument evaluation involved a sophisticated thinking process, even if it was the "evaluation of claim alone." As the interview went along, Eli mentioned that he had also examined the evidence and justification in Argument A, yet didn't write it down in the worksheet. He explained that "Argument A did not really explain well" and lacked critical evidence, which was that "air has weight" and Argument A "should clearly say" it (Turn 1-6).

The second example is Oliver's argument evaluation. Oliver thought Argument B was the most convincing; he wrote on the worksheet "because if I have an inflated soccer ball, it is lighter than the deflated one." During the interview, he explained how he thought.

2-1 Oliver: Yes, B is the most convincing, and so B is correct.
2-2 Interviewer: How do you know that? What was your thinking process of making that judgment?
2-3 Oliver: I have the same experience. I play soccer … the moment I read his [Argument B's] evidence, I knew it is correct! Believe me, it is correct! So it is the most convincing.
2-4 Interviewer: You examined the evidence to determine whether it is convincing or not?
2-5 Oliver: Yes, I examined the evidence. That is my thinking process, I think the evidence is convincing, so the entire argument is convincing. Like, the evidence is the most important … Because the Evidence B used is correct, so it is most convincing … Because B is the most convincing argument, so it is correct. I agree with it.

According to what Oliver shared, when he read Argument B's evidence, he knew the evidence was correct and convincing to him. In particular, he referred back to his own experience of playing soccer and realized that he had had "the same experience" of finding it difficult to kick a flat soccer ball far away (Turn 2-3). Then, he elaborated that "because the Evidence B used is convincing, so the entire argument is convincing" (Turn 2-5). He also demonstrated his epistemic belief that

"evidence is the most important" to make a claim (Turn 2-5). Finally, because he thought Argument B was "the most convincing," he concluded that "B is correct" (turn 2-5). In other words, for Oliver, the process of his argument evaluation was: he started by examining the evidence; then he chose the argument with the most persuasive evidence as the most convincing one; and finally, he determined that the most convincing argument was correct. He thought the most convincing argument answered the question correctly, rather than choosing the correct argument as the most convincing.

Through comparing these two examples, it was found that both Eli and Oliver thought the argument that made the "correct" claim (they thought) was the most convincing, and in both cases, their epistemic understandings were involved in their argument evaluation. However, their thinking processes were different from each other. For example, Eli thought the argument was convincing because it was correct; however, Oliver thought that because it was convincing, then the argument was correct. SRIs with them showed that the processes of their argument evaluation were more sophisticated than what was represented in their written answers.

Evaluating Arguments With the Same Claim

The second part of the argument evaluation task asked students to choose the more convincing argument from two that made the same claim but with different evidence and reasoning. To this question (i.e., Which one is more convincing?), students also provided different answers. Two students thought Argument A was more convincing, and the other 14 students chose Argument E as the more convincing one. Even if they made different choices, the written responses and interviews with both groups of students revealed that they all examined the relationship between claim and evidence to make their judgments. When students compared arguments with the same claim, it was evident that they tended to focus on the sufficiency and relevance of evidence and the quality of reasoning. For example, students examined whether the evidence was relevant and enough to make the claim and whether the reasoning was clear to justify the claim's validity. These trends were demonstrated in both their written responses and what they verbally shared in the SRIs. For example, in both her written answers and the SRI with her, Lisa indicated that she thought Argument E was more convincing than A "because E said more and had more evidence than A." David also thought Argument E was more convincing, and his written reasons were "because E has more evidence and uses more convincing words, such as 'same amount.'" In the SRI with him, David confirmed that he made his choice by comparing the evidence and reasoning in both arguments:

> I was thinking, since they have the same idea, the inflated soccer ball is heavier, so I might need to focus on what evidence they had and how they explained … So, I compared their evidence. E has more evidence, which is good. And, yes, it explains better with more convincing words, like "same amount." So overall, E has more evidence and explained better, so E is more convincing than A … That was my thinking process.

Another student, Ming, chose Argument A as the more convincing one. He wrote in his handout that "A is more convincing, because A is more concise, easier to understand. E says more, but too much." In the later SRI, Ming elaborated as follows:

> My thinking process was, I think I need to compare their differences to see which was more convincing … Even if E said more, but don't you think it is too much? And it said the inflated ball has more air, that is confusing to me … I don't think so. The evidence [of E] is not good, so it [argument E] is not convincing.

These examples illustrated how students evaluated arguments that made the same claim. After realizing the two arguments shared the same claim, students examined the evidence and reasoning to make their judgment about which one was the stronger argument. This trend was evident in both their written responses and SRIs with them, no matter what specific argument they chose as the more convincing one. Yet, compared with students' written answers, SRIs provided more sufficient and more detailed information, which was helpful for understanding students' argument evaluation more comprehensively.

The Influence of SRI on Students' Thinking

It has been suggested in the literature that interviews aiming to access students' thinking would influence their thinking as well, because the probing and questioning in these interviews would stimulate students' reflections on their own thinking and awareness (Thomas, 2013). In this study, this influence was also noticed. It was found the SRIs with students affected their thinking and how they made their judgments of argument evaluation. Some episodes that illustrated this notion were discerned. What follows is one of them.

After students completed their argument evaluation tasks, Jaden volunteered to participate in the SRI and share his thinking processes of doing the task. When he was explaining how he decided Argument E was more convincing than Argument A, we had the following conversation.

> 3-1 Jaden: Student E is this long (showing with his fingers that Student E has many lines of words), while Student A is only this long (showing with his finger that Student A has fewer lines of words), so it is obvious that E is more convincing.
> 3-2 Interviewer: This is the thinking process of your decision making at that time, right?
> 3-3 Jaden: Yes.
> 3-4 Interviewer: So, your way of thinking is to examine which one has more words. You think the more words the more convincing, right? If I said more, a lot a lot and a lot, then it will make whatever I said more convincing, you mean this?
> 3-5 Jaden: Yes (being silent for a few seconds). No, wait, hmm. No, let me think (being silent for many seconds).
> …
> 3-6 Jaden: No. I think E is more convincing, because, yes, she said more, and that helps. But she also gave more evidence and, like, explained better. She did not, like, just blahblahblah, she had evidence.
> 3-7 Interviewer: Then, when you think back now, did you make your choice only because E said more than A?
> 3-8 Jaden: No, I think … No. I remember that when I read Student C, I thought he was just like blahblahblah, said a lot a lot, but nothing really make sense. But not Student E.
> 3-9 Interviewer: So, you mean that you examined whether Student E was just blahblahblah or her evidence made sense, right?
> 3-10 Jaden: Yes, I think I examined her evidence. So, can I change my answer on this?

Initially, Jaden said he thought Argument E was more convincing because E was longer than A (Turn 3-1). When the interviewer further questioned his thinking process by repeating and

rephrasing his words (Turn 3-4), he was challenged and revisited his thinking process back then (Turn 3-5). With the questions and cues, he engaged in active reflection on his own thinking process. Then, he realized that when he made his decision, he also examined the quality of E's words, such as whether E used supportive evidence or whether the cited evidence made sense (Turns 3-8, 3-10). In this way, during the SRI and with the interview questions about his thinking process, Jaden reflected on and became aware of his thought process.

With the SRIs, it was found that some students not only were aware of their thought process but were also capable of describing and discussing it with others, while other students might be not aware of their own thought processes. For those students who did not have the awareness, their retrospectively self-reported thought process cannot reflect how they actually thought. With the researcher's explicit questioning or prompting, they actively reflected on and recalled more accurately their thought process back then. In this way, the interviewer's questioning and prompting stimulated their active reflection on their own thinking. With these identified episodes, the researcher acknowledges that SRI as a data collection method to gain relevant information about students' thinking process in this study was not neutral; instead, it influenced students' thinking.

Limitation and Delimitation of This Study

The gender inequality in the student participants is a limitation of this study, especially among students who participated in the SRIs (seven boys and one girl). The gender composition of the class contributed to some extent to this inequality. Among the 19 student participants, only four were girls. Moreover, compared with the boys, these girls were quieter, and only one of them volunteered to participate in interviews. For these reasons, there was gender inequality in the sample.

This study explored how students performed the argument evaluation task, with a particular emphasis on how they thought in order to make their judgment. The correctness of their answers to the argument evaluation questions was not the focus of this paper, such as whether the choices they made and their reasons were correct or not. Questions such as "Why did some students not apply their science knowledge in the task to make the right choice?" and "How might teachers scaffold students' conceptual change?" are of pedagogical significance. However, this study cannot answer these questions because they are beyond the scope of this case study.

The argument evaluation task that students completed in this study was about a well-structured scientific problem (i.e., whether air has weight). It has been suggested in the literature that how students perform argumentation to explore socioscientific issues (SSI), which are usually ill-structured, open-ended, and value-involved, is different from how they solve well-structured scientific problems through argumentation (Nielsen, 2012; Jonassen, 2000). Therefore, it is acknowledged that findings from this study might not be able to explain how students evaluated arguments in the context of SSI. Future research with a focus on SSI is needed to help us understand students' argument evaluation comprehensively.

Conclusions, Implications, and Further Considerations

This study extended prior research on students' argument evaluation by exploring a new research method—stimulated recall interview (SRI)—to investigate how students compare and evaluate arguments constructed by others. Together with a written task, SRI provided informative data for

the researcher to understand the process of students' argument evaluation, including not only their observable performance but also their inner awareness and thinking process of making their judgments. Regarding students' argument evaluation, the findings of this study suggest that students' thinking processes are usually more sophisticated than what is represented in their written work. Most of the existing research takes students' written responses as the main data source to explore students' argument evaluation (Kuhn et al., 2013; Ryu & Sandoval, 2012). Therefore, this study considered that only examining students' written responses to an argument evaluation task might be not sufficient to fully explain the complexity of their argument evaluation. For example, when students in this study evaluated various arguments, especially arguments making different claims, in addition to examining the correctness of the claims, they also examined the evidence cited in the arguments. Yet, what they wrote down on their worksheets was mainly about the claims, such as whether the claim in the argument was correct and whether they agreed with it. Many examples discerned in this study indicate that students tended to simplify the process of their argument evaluation when they were asked to write it down. Moreover, even if some students only examined the correctness of claims while they were evaluating various arguments, as shown in this study, their thinking processes might be not as simple as had been assumed. Therefore, more consideration is required of whether "examining claims alone" is the indicator of lower competency in argument evaluation.

To better understand students' argument evaluation, this study suggests considering their inner thinking processes. Regarding the methodological decision to investigate students' thinking, this study suggests that SRI is a feasible and fruitful method. This kind of interview has the potential to provide researchers informative data because it affords researchers opportunities to learn how students thought in order to make their judgments and rationales for their decisions, which they usually did not explain in the written tasks. Combining SRI with written tasks is one of the possible ways to understand students' argument evaluation holistically.

Together with the confirmation that SRI could be a helpful technique to learn about students' argument evaluation, other noteworthy issues about this particular kind of interview were also revealed. The first is that SRI is not a neutral data collection method. Instead, as a form of social interaction (Deppermann, 2013), it influences students' thinking and doing. Questions and prompts within the interviews have the potential to stimulate students' active reflection on their own thinking, which might not happen to some students without their being asked these questions. This echoes what Thomas (2012) explains: that any research into metacognition or thinking about thinking (Flavell, 1979) is itself a form of intervention in students' thinking and behavior. Participating SRIs provides students opportunities to explicitly explain their thinking processes. When students' thinking is explicit in social context, social interactions such as the conversations with interviewer during the SRIs influences how they would think and do. This suggests that classroom teaching could incorporate such practice to support student learning through making their thinking visible. When classroom teaching allows and encourages students to share not only their ideas about topics under investigation but also their thinking processes, teachers have opportunities to understand and facilitate students' ways of thinking, which goes beyond the conceptual domain (Duschl, 2008).

Another issue that presents a challenge for SRIs and other self-report interviews is social desirability. Participants' responses may be affected by their own expectations and the perceived expectations of others, such as the interviewer (Thorndike, 2005). To minimize the influence of social desirability, it is very important to have research participants feel safe and to ensure that they know their thoughts are valued instead of being judged. In this study, the researcher ensured

before each SRI that students knew that any idea or thought they shared during the interview would be valued instead of judged. In this way, the researcher managed to build a trust-based relationship with the students, which is critical for collecting informative data through the self-report interview (Groenewald & Bhana, 2015), in which "the level of information revealed is controlled by the group members being investigated" (Merriam, 2016, p. 145). Researchers who want to try this method in their research should consider these issues when they design their research.

References

Akturk, A. O., & Sahin, I. (2011). Literature review on metacognition and its measurement. *Procedia–Social and Behavioral Sciences, 15*, 3731–3736. https://doi.org/10.1016/j.sbspro.2011.04.364

Anderson, D., Nashon, S. M., & Thomas, G. P. (2009). Evolution of research methods for probing and understanding metacognition. *Research in Science Education, 39*(2), 181–195. https://doi.org/10.1007/s11165-007-9078-1

Bernard, R. C. (1994). *Research methods in anthropology: Qualitative and quantitative approaches* (2nd ed.). AltaMira Press.

Bogdan, R. C., & Biklen, S. K. (2007). *Qualitative research for education: An introduction to theory and methods* (5th ed.). Allyn & Bacon.

Branch, J. L. (2006). Using think alouds, think afters, and think togethers to research adolescents' inquiry experiences. *Alberta Journal of Education Research, 52*(3), 148–159. https://doi.org/10.11575/ajer.v52i3.55153

Breakstone, J., McGrew, S., Smith, M., Ortega, T., & Wineburg, S. (2018). Why we need a new approach to teaching digital literacy. *Phi Delta Kappan, 99*(6), 27–32. https://doi.org/10.1177/0031721718762419

Cavagnetto, A. R. (2010). Argument to foster scientific literacy: A review of argument interventions in K–12 science contexts. *Review of Educational Research, 80*(3), 336–371. https://doi/10.3102/0034654310376953

Chen, Y-C. (2020). Dialogic pathways to manage uncertainty for productive engagement in scientific argumentation: A longitudinal case study grounded in an ethnographic perspective. *Science & Education, 29*(2), 331–375. https://doi.org/10.1007/s11191-020-00111-z

Chen, Y-C., Hand, B., & Park, S. (2016). Examining elementary students' development of oral and written argumentation practices through argument-based inquiry. *Science & Education, 25*(3/4), 277–320. https://doi.org/10.1007/s11191-016-9811-0

Cheng, C.-H., & Yang, F.-Y. (2022). Analyzing visual attention during TAP learning and the effect of epistemic beliefs on the understanding of argument components. *International Journal of Science Education, 44*(8), 1336–1355. https://doi.org/10.1080/09500693.2022.2076950

Choi, A., Notebaert, A., Diaz, J., & Hand, B. (2010). Examining arguments generated by year 5, 7, and 10 students in science classrooms. *Research in Science Education, 40*(2), 149–169. https://doi.org/10.1007/s11165-008-9105-x

Deppermann, A. (2013). Interviews as text vs. interviews as social interaction. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 14*(3). http://doi.org/10.17169/fqs-14.3.2064

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

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education, 38*, 39–72. https://doi.org/10.1080/03057260208560187

Eugenio-Gozalbo, M., Zuazagoitia, D., Ruiz-González, A., Corrochano, D., Hurtado-Soler, A., & Talavera, M. (2022). Implementing citizen science programmes in the context of university gardens to promote pre-service teachers' scientific literacy: A study case on soil. *International Journal of Science*

Education, 44(10), 1619–1638. https://doi.org/10.1080/09500693.2022.2088877
Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906
Groenewald, C., & Bhana, A. (2015). Using the lifegrid in qualitative interviews with parents and substance abusing adolescents. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 16*(3). https://doi.org/10.17169/fqs-16.3.2401
Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). SAGE.
Guilfoyle, L., & Erduran, S. (2021) Recalibrating the evolution versus creationism debate for student learning: towards students' evaluation of evidence in an argumentation task. *International Journal of Science Education, 43*(18), 2974–2995. https://doi.org/10.1080/09500693.2021.2004330
Hosbein, K. N., Lower, M. A., & Walker, J. P. (2021). Tracking student argumentation skills across general chemistry through argument-driven inquiry using the assessment of scientific argumentation in the classroom observation protocol. *Journal of Chemical Education, 98*(6), 1875–1887. https://doi.org/10.1021/acs.jchemed.0c01225
Iordanou, K., Kendeou, P., & Beker, K. (2016). Argumentative reasoning. In J. A. Greene, W. A. Sandoval & I. Bråten (Eds.), *Handbook of epistemic cognition* (pp. 39–53). Routledge.
Jin, Q., & Kim, M. (2021). Supporting elementary students' scientific argumentation with argument-focused metacognitive scaffolds (AMSs). *International Journal of Science Education, 43*(12), 1984–2006. https://doi.org/10.1080/09500693.2021.1947542
Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63–85. https://doi.org/10.1007/BF02300500
Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *Journal of the Learning Sciences, 4*(1), 39–103. https://doi.org/10.1207/s15327809jls0401_2
Kim, M. (2016). Children's reasoning as collective social action through problem solving in grade 2/3 science classrooms. *International Journal of Science Education, 38*(1), 51–72. https://doi.org/10.1080/09500693.2015.1125559
Kim, M., & Roth, W-M. (2014). Argumentation as/in/for dialogical relation: A case study from elementary school science. *Pedagogies, 9*(4), 300–321. https://doi.org/10.1080/1554480X.2014.955498
Kuhn, D. (2019). Critical thinking as discourse. *Human Development, 62*, 146–164. https://doi.org/10.1159/000500171
Kuhn, D., Zillmer, N., Crowell, A., & Zavala, J. (2013). Developing norms of argumentation: Metacognitive, epistemological, and social dimensions of developing argumentive competence. *Cognition and Instruction, 31*(4), 456–496. https://doi.org/10.1080/07370008.2013.830618
Lytzerinou, E., & Iordanou, K. (2020). Teachers' ability to construct arguments, but not their perceived self-efficacy of teaching, predicts their ability to evaluate arguments. *International Journal of Science Education, 42*(4), 617–634. https://doi.org/10.1080/09500693.2020.1722864
Malva, L., Leijen, Ä., & Arcidiacono, F. (2021). Identifying teachers' general pedagogical knowledge: A video stimulated recall study. *Educational Studies*. https://doi.org/10.1080/03055698.2021.1873738
McNeill, K. L., González-Howard, M., Katsh-Singer, R., & Loper, S. (2016). Pedagogical content knowledge of argumentation: Using classroom contexts to assess high-quality PCK rather than pseudoargumentation. *Journal of Research in Science Teaching, 53*(2), 261–290. https://doi.org/10.1002/tea.21252
Merriam, S. B. (2016). *Qualitative research: A guide to design and implementation*. Jossey-Bass.
Münchow, H., Richter, T., von der Mühlen, S., & Schmid, S. (2019). The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, nomological network, and relevance for academic success at the university. *British Journal of Educational Psychology, 89*(3), 501–523. https://doi.org/10.1111/bjep.12298
Nielsen, J. A. (2012). Science in discussions: An analysis of the use of science content in socioscientific

discussions. *Science Education, 96*(3), 428–456. https://doi.org/10.1002/sce.21001
Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 44*(10), 994–1020. https://doi.org/10.1002/tea.20035
Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). SAGE.
Rowe, V. C. (2009). Using video-stimulated recall as a basis for interviews: Some experiences from the field. *Music Education Research, 11*(4), 425–437. https://doi.org/10.1080/14613800903390766
Ryu, S., & Sandoval, W. A (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education, 96*(3), 488–526. https://doi.org/10.1002/sce.21006
Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future direction. *Science Education, 92*(3), 447–472. https://doi.org/10.1002/sce.20276
Sampson, V., & Clark, D. B. (2009). The impact of collaboration on the outcomes of scientific argumentation. *Science Education, 93*(3), 448–484. https://doi.org/10.1002/sce.20306
Sampson, V., Grooms, J., & Walker, J. P. (2011). Argument-driven inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education, 95*(2), 217–257. https://doi.org/10.1002/sce.20421
Settlage, J., & Southerland, S. A. (2019). Epistemic tools for science classrooms: The continual need to accommodate and adapt. *Science Education, 103*(4), 1112–1119. https://doi.org/10.1002/sce.21510
Shaw, V. F. (1996). The cognitive process in informal reasoning. *Thinking & Reasoning, 2*(1), 51–80. https://doi.org/10.1080/135467896394564
Takao, A. & Kelly, G. (2003). Assessment of evidence in university students' scientific writing. *Science & Education, 12*, 341–363. https://doi.org/10.1023/A:1024450509847
Thomas, G. P. (2012). Metacognition in science education: Past, present, and future considerations. In B. Fraser, K. Tobin & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 131–144). Springer.
Thomas, G. P. (2013). The interview as a metacognitive experience for students: Implications for practice in research and teaching. *Alberta Science Education Journal, 43*(1), 4–11.
Thomas, G. P., & McRobbie, C. J. (2001). Using a metaphor for learning to improve students' metacognition in the chemistry classroom. *Journal of Research in Science Teaching, 38*(2), 222–259. https://doi.org/10.1002/1098-2736(200102)38:2<222::AID-TEA1004>3.0.CO;2-S
Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Pearson Prentice Hall.
Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.
Veenman, M. V. J. (2005). The assessment of metacognitive skills: What can be learned from multi-method designs? In C. Artelt & B. Moschner (Eds.), *Lernstrategien und metakognition: Implikationen für forschung und praxis [Learner strategies and metacognition: Implications for research and praxis]* (pp. 75–97). Waxmann.
Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0

Qingna Jin is a Researcher at the Faculty of Education, University of Alberta. Her research interests include students' dialogical argumentation in school science and STEM learning contexts, underrepresented students in science classrooms, and children's metacognition in learning. https://orcid.org/0000-0003-2580-4964

Correspondence concerning this article should be addressed to Qingna Jin, 551 Education South, Department of Elementary Education, University of Alberta, 11210-87 Avenue, Edmonton, AB, Canada T6G 2G5. Email: qingna@ualberta.ca