European Journal of Science and Mathematics Education, 2023, **11**(4), 615-634
ISSN: 2301-251X (Online)

OPEN ACCESS

Research Article

What do we really know about students' written arguments? Evaluating written argumentation skills

Maria Evagorou [1*]
0000-0002-1299-946X

Elena Papanastasiou [1]
0000-0001-5634-8450

Maria Vrikki [1]
0000-0002-2748-4418

[1] Department of Education, School of Education, University of Nicosia, 2417, Nicosia, CYPRUS
[*] Corresponding author: evagorou.m@unic.ac.cy

Citation: Evagorou, M., Papanastasiou, E., & Vrikki, M. (2023). What do we really know about students' written arguments? Evaluating written argumentation skills. *European Journal of Science and Mathematics Education, 11*(4), 615-634. https://doi.org/10.30935/scimath/13284

ARTICLE INFO

Received: 31 Oct 2022

Accepted: 24 Apr 2023

ABSTRACT

The purpose of this study was to explore the different sub-skills of students' written arguments (i.e., writing an argument, choosing a convincing argument) that might exist, and the content dependency of arguments. This paper presents two written argumentation tools that were designed for 11-14 year-old students, and the main outcomes from applying the tools to evaluate the written arguments of 246 students. The analysis of the data implies that choosing a convincing argument is a different kind of skill than any of the other three aspects of argumentation that were evaluated in these tests; that argumentation is content specific, and that argument construction is easier when the students' have knowledge of the topic, regardless of whether this is a scientific or an everyday life topic. A main contribution in this study is that we have identified the degree of complexity for all four sub-skills that were included in the test. By identifying that writing an argument is a more difficult skill to acquire, or that students are not acquainted with it, it can help educators to design better scaffolding structures to support students when writing counterarguments. Research implications arising from the findings include exploring in detail how students choose to agree or disagree with given claims in different situations – for example exploring the difference in agreeing with media claims on socioscientific issues as opposed to scientific claims in the science classroom. Implications for teaching include using different teaching approaches for scientific and everyday argumentation.

Keywords: argument, science education, written argument, scientific reasoning

INTRODUCTION

People are called to make decisions in their everyday lives that require scientific knowledge (Allchin, 2022; Sharon & Baram-Tsabari, 2020) and argumentation. Recent political and social developments across the globe, such as vaccination against COVID19 and climate change, have highlighted the need for people to be scientifically literate to be in a position to make such important decisions. However, the lack of training in research methodology perpetuates the inability of individuals to evaluate claims that are made online and elsewhere (Papanastasiou, 2019). As a result, efforts to sway groups of people towards a specific direction have given rise to waves of misinformation and fake news in an unprecedented way; this makes the need for people to develop their scientific literacy and argumentation even more pertinent. Scientific literacy is a complex construct with several components, such as identifying misinformation, understanding the practice of science and judging scientific expertise (Sharon & Baram-Tsabari, 2020, p. 875-876). Scientific literacy involves, amongst others, being able to read and comprehend scientific articles and articles about science, and engage in oral and written exchanges about the validity of the conclusions. Furthermore, scientific literacy involves reasoning about science, considering counter-arguments and being able to refute arguments and to re-evaluate one's position (Voss & Means, 1991). Argumentation is therefore an important component of scientific literacy because it promotes both critical thinking skills, understanding of science and the connection to everyday life (Allchin, 2023).

Argumentation has been a prominent field of study, especially within the context of science education research during the past decades (Erduran, 2022; Erduran et al., 2015; Lee et al., 2009) and studies have ranged from the analysis of students' written and oral arguments to teaching methods promoting argumentation. Specifically, researchers have analysed interviews in which people engaged in oral argumentation (i.e., Kuhn, 1991), students' written arguments from science lessons (i.e., Jimenez-Aleixandre & Pereiro-Munoz, 2002; McNeill & Pimentel, 2009; Rapanta & Christodoulou, 2022), students' artefacts created during the instruction (Sampson & Clark, 2008), and written essays or texts (Erduran et al., 2015; Kelly & Takao, 2002; Macagno, 2016; Osborne et al., 2004). The aforementioned studies have identified a number of difficulties that students face with written and oral argumentation and have proposed instructional approaches that can help enhance both written and oral argumentation in the classroom and tried to identify learning progressions for argumentation (Osborne et al., 2016). Based on the findings from previous studies, it is apparent that students' difficulties with both written and oral argument have been well researched and identified, especially in science education (Kelly & Takao, 2002; Osborne et al., 2004). However, most of the studies on written argumentation place an emphasis on only one aspect - writing an argument mostly in the form of a written essay. However, sub-skills of written argumentation as for example deciding which is a more sophisticated argument or counter-argument, and how students argue in different contexts is relatively unexplored (Macagno, 2016; Rodriquez-Mora et al., 2022). Therefore, what is identified as missing from research in argumentation in science education is an exploration of the various sub-skills associated with written argument. More specifically: (a) students' ability to evaluate written arguments (Bravo-Torija & Jimenez-Aleixandre, 2018), (b) students' ability to respond to an argument by writing their own counter-arguments (i.e., Osborne et al., 2016), (c) whether written argumentation is a context-specific skill or not, and (d) whether writing an argument or a counter-argument are at the same difficulty level compared to deciding which is the best argument from a given list. Studies from fields outside science education are informative to some of the issues, i.e., content-specificity based on the scenario (Ray et al, 2018), but not conclusive, and more recent studies from science education provide insight on the impact of the structure of the question on students final written arguments (i.e., Rodriquez-Mora, Cebria-Robles & Blanco Lopez, 2022).

Based on the aforementioned gap, the purpose of this study was threefold: (a) to explore if there is a difference in the level of complexity in students written argumentation sub-skills (i.e., choosing a convincing argument, choosing a convincing counter-argument, writing an argument, writing a convincing counter-argument), (b) to explore if written argumentation is content specific, and (c) to explore the structure of students' written arguments and counterarguments. More specifically the research questions guiding this study are:

(a) To what extend is there a difference in the level of complexity in students' written argumentation sub-skills (i.e., choosing a convincing argument, choosing a convincing counter-argument, writing an argument, writing a convincing counter-argument)?

(b) To what extend is written argumentation content specific?

(c) What is the structure of students' written arguments and counterarguments?

THEORETICAL FRAMEWORK AND LITERATURE REVIEW

Argumentation, broadly speaking, refers to the ways that evidence is used to persuade a critic of the merits or lack of a standpoint or position (van Eemeren & Grootendorst, 2003). It is a specific form of talk that enable students to communicate in the classroom in the ways similar to their everyday lives, and help them view science as an epistemological and social process in which knowledge claims are generated, adapted, reorganized, and, at times, abandoned (Lawson, 2003; Lederman, 1992). Even though argumentation is a specific form of talk, it is also viewed as a written activity aimed at justifying or defending a standpoint for an audience (van Eemeren et al., 1996). Arguments in this study are defined as the written set of claims, data, warrants, and backings that contribute to the content of an argument. In Toulmin's Argument Pattern (TAP), the essential elements are claims, data, warrants and backings. Toulmin defines data as 'the facts we appeal to as a foundation for the claim' and warrants 'general hypothetical statements, which can act as bridges' (p. 97-98). According to TAP, data are the facts that those involved in the argument appeal to in support of their claim. A claim is the conclusion whose merits are to be established. Warrants are the reasons that are used to justify the connections between the data and the conclusion, and backings are the basic assumptions that provide the justification for particular warrants. Additionally, in more complex arguments, Toulmin identifies two more features in his framework; the qualifiers that specify the conditions under which the claim is true – and rebuttals – which specify the conditions in which the claim may not be true. All studies using Toulmin's framework have focused on the structural issues, and provide information on how students structure their arguments, and the kind of difficulties they have (e.g., Bell & Linn, 2000; Osborne et al., 2004), and providing in that way guidelines for designing effective argumentation learning environments. The main criticism of Toulmin's framework is that it is not easy to distinguish between claims, data, warrants and qualifiers (Sampson & Clark, 2008), because the decision of what counts as data, warrants and claims depends on what was said exactly before that in the dialogue, and to what that refers. Hence, either the researcher has to make an inference (Erduran et al., 2004), or the terms have to be better defined, using indicating words to identify when something is a claim, a warrant or a rebuttal.

A second criticism of Toulmin's framework is that it is a domain general framework, which only refers to the structure of the argument (that is the presence or not of claims, warrants, rebuttals) and does not evaluate the content (Sampson & Clark, 2008). So, even though an argument might be considered high quality in terms of structure, the accuracy of the content might not be relevant, and must be supplemented by an additional analysis of the content. To address the methodological issue of deciding about the quality of the arguments, Erduran et al. (2004) devised five argument levels to 'measure' or explain the quality of argument and argumentation, especially as a measure of interactive discourse, since the main identifier of quality in their levels is the presence or not of rebuttals (Erduran, 2008). These levels are based theoretically on Toulmin's framework and are informed from empirical evidence on how young students construct arguments (e.g., Osborne et al., 2004). The authors suggest the following levels of argumentation:

- Level 1: arguments that are a simple claim versus a counter-claim or a claim versus a claim.
- Level 2: consist of a claim versus a claim with either data, warrants, or backings but which does not possess any rebuttals.
- Level 3: consists of a series of claims or counter-claims with either data, warrants, or backings with the occasional weak rebuttal.
- Level 4: arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counter-claims.
- Level 5: an extended argument with more than one rebuttal (Erduran et al., 2004)

In designing and evaluating the data collection tool for this study, a modified version of TAP as proposed by Erduran et al. (2004) was used as a guide. The value of this modified version of Toulmin's framework lies in the fact that it enables an identification of the level, or what might be termed the quality of argument, and can be used to evaluate written arguments, even though the presence of rebuttals in written arguments should not be expected to be as frequent. The choice of this framework is based mainly on the fact that is has been previously applied for the analysis of arguments and argumentation for a similar age group as the one in the current study (Osborne et al., 2004), it has been widely used by science education researchers (e.g., Osborne et al., 2004).

Finally, the theoretical framework of this work has also been influenced by the idea that argumentation is 'a fundamental tool of reasoning' (Voss & Means, 1991, p. 337) that can be applied to different contexts (Mason & Scirica, 2006). Therefore, the tool used in this study was designed in a way to reflect this idea.

Students' Difficulties in Written Argument Construction

When it comes to written arguments, studies in the field of science education have revealed amongst others that students struggle with scientific explanation even though they are good at supporting ideas, challenging and counter-challenging during everyday conversations (i.e., Ferretti & Graham, 2019; Kuhn, 1991), they tend to use inappropriate reasoning (i.e., Zeidler, 1997), or irrelevant data (i.e., McNeill & Krajcik, 2008; Sandoval & Millwood, 2005), they distort or ignore evidence in an effort to support their own conceptions (i.e., Sampson & Clark, 2011), and they focus on their claims without necessarily justifying them (i.e., Jiménez-Aleixandre et al., 2000). In a study with secondary school biology students which focused on students' coordination of evidence with their causal claim, students distinguished claims from data, but students' references to data in their written explanations failed to interpret meaning of those data, since they believed that data spoke for themselves and there was no need to explain them (Sandoval & Millwood, 2005). Additionally, the same researchers found that the students reacted differently in terms of coordinating evidence with claims in their written arguments for the two different topics that were presented to them. Bell and Linn (2000) analysed written arguments of high school students and their analysis helped them relate the structure of the argument to conceptual understanding, but not the structure to the quality. Kelly and colleagues (Kelly & Takao, 2002) also explored undergraduate students' written arguments during an oceanography lesson. Their findings suggest that some students can write essays that are based on well-presented and supported arguments, but some others provide poorly evidenced arguments which can either be based on vague reference to supporting data, be based on a number of evidence without managing to create an argument based on those, or the written argument is based on minimal data. Finally, Rodríguez-Mora et al. (2022) found that the most difficult element for students in developing a written argument was using evidence, followed by justifications and finally conclusions.

Students' Difficulties in Evaluating Arguments

Other than writing arguments, another aspect of argumentation is evaluating written arguments (Macagno, 2016). Studies in the area of evaluating written arguments are sparse, especially studies with middle school students. Norris et al. (2003) for example presented 12th grade students with a media report and a series of statements which they were asked to evaluate. Their findings suggest that fewer than half of the students were able to interpret causal written statements, almost half understood evidence statements as conclusions, and 90 per cent recognized observations as such. In a similar study Norris and Phillips (2003) presented 380 university students with five media reports and asked them to answer questions about how they interpret the reports, and to make judgments about the certainty, status, and role of the statements identified in the report. The findings of this study suggest that the university students confused cause and correlation, and they had difficulties distinguishing explanation of phenomena from the phenomena themselves. Finally, Gleim et al. (2010) investigated how fifty middle and high school students evaluate science related claims found in popular media and what characteristics of the arguments in the media the students find more persuasive. Their findings suggested that the students needed more proof/data, and wanted more scientists to talk about it before agreeing with the arguments presented. Summarizing, students find it difficult to interpret written causal statements, and identify data, which supports the statements, presented in the text, they believe that more data makes a stronger argument, and that scientists agreeing with a statement makes an argument stronger.

Content Specificity of Argumentation

Studies exploring the issue of content specificity of informal reasoning (i.e., critical thinking and argumentation) thus far are inconclusive. For example, in a study with people of different ages and three everyday scenarios, Kuhn concluded that people reason better in subjects in which they have personal knowledge (Kuhn, 1991). Along the same lines, a study with 8th graders, in the context of a socioscientific issue, showed that better prior knowledge helped the students construct better rebuttals, but this association was not clear for the construction of arguments and counter-arguments (Mason & Scirica, 2006). Additionally, research indicates that poor performance in argumentation is associated with lack of scientific knowledge (Jiménez-Aleixandre et al., 2000; Sandoval & Reiser, 2004). van Aufschnaiter et al. (2008) argue that "even scientists may not be able to engage in high level argumentation when confronted with an unfamiliar task. Hence to promote high level argumentation and students' understanding of scientific concepts, it is essential to consider both the relevance of students' prior experience and the complexity of the tasks" (p. 24). On the other hand, Voss and Means (1991) showed that the ability and knowledge of university students in economics did not yield any difference in students' reasoning. Additionally, the same researchers describe a case study

Table 1. Description of parts of assessment test, corresponding levels of arguments and scores

	Description of aspects of argument	Numerical Score
Level 5	Data, warrant, and rebuttal (most convincing argument)	4
Level 4	Warrant and data	3
Level 3	Data only	2
Level 2	Warrant only	1
Level 1	Claim only, no response, irrelevant	0

of a student with good reasoning skills, but no prior knowledge of the subject, trying to write an essay on earthquakes using a think aloud protocol. The student generated and evaluated reasons and engaged in argumentation, without any prior knowledge on the domain that was presented to him (Voss & Means, 1991). Even though the studies described above are non-conclusive with respect to the role of domain knowledge, most of the studies agree that schooling (or age) does not improve reasoning (Kuhn, 1991; Perkins & Salomon, 1989), and that the best predictors for reasoning are the person's abilities (Kuhn, 1991; Perkins, 1993), and epistemological dispositions (Mason & Scirica, 2006). Despite the contradicting findings in previous studies regarding the content specificity of reasoning skills, this study assumes a situated cognition stance, and therefore the assessment tool in this study was designed based on the assumption that knowledge of the domain might have an effect on students' argumentation.

METHOD

In order to explore if there is a difference in the level of complexity in students written argumentation sub-skills an assessment tool was designed to measure 11–14-year-old students' ability to write arguments and counter-arguments, and their ability to evaluate written arguments and counter-arguments in different contexts.

Assessment Tool

The theoretical framework underlying the design of the assessment test was that of Toulmin's view of the elements of an argument. According to Toulmin's framework the essential elements of an argument are claims, data, warrants and backings. Counter-arguments are also important, especially in a dialogue. The decision to ask students to state which is a constructive argument and which is a constructive counter-argument, from a given list, lies on the fact that this is considered an important skill in everyday life (Macagno, 2016). **Table 1** presents the levels of the argument based on a modified version of Toulmin's framework (Erduran et al., 2004) which provides levels of complexity for the arguments. For example, a Level 5 argument is better quality than a Level 1 argument. For the purposes of the statistical analyses the levels were translated in scores as shown in the last column of **Table 1**.

During the initial phase of the design of the tool, various questions were designed to assess four different aspects of argumentation, namely: (a) deciding what a convincing argument is; (b) deciding what a convincing counter-argument is; (c) constructing convincing arguments and; (d) constructing convincing counter-arguments. The choice of these four parts is based on the theoretical framework of this study, in which writing and reading/evaluating arguments are considered important aspects of scientific literacy. The questions in the assessment tool are either based on evaluation questions from previous studies, for example the IDEAS project (Osborne et. al., 2004) and TIMSS (Garratt et al.,1999), or were specially designed by the first author for the purposes of this study.

Specifically, in order to measure the construct of written argumentation, eight items were developed consisting of two items for each sub-skill (choosing a convincing argument, choosing a convincing counter-argument; constructing a written argument, constructing a written counter-argument,). The items were pilot tested with 21, 12–13-year-old students coming from an urban UK school. Classroom observations and an initial analysis suggested that the questionnaire was too long for the students to complete, and hence it was decided to create the two shorter versions (Tool A and Tool B). One of the initial assumptions in this study is that both tools measure the same construct: written argument. It is important to note that the first two questions in both tests were multiple-choice questions, and the students were not taught the topics of the questions. For these questions, five specific answers were provided, with each answer representing one of

Table 2. Overview of structure and content of the two questionnaires

	Choosing a convincing argument	Choosing a convincing counter-argument	Constructing an argument	Constructing a counter argument
Tool A	Q1a: Is current used up in a simple electric circuit? Choose the appropriate answer from the list and justify your choice.	Q2a: Does light travel from our eyes to objects? Choose the appropriate answer from the list and justify your choice.	Q3a: Using chemicals to kill mosquitoes?	Q4b: Wind energy vs. nuclear factories
Tool B	Q1b: Explain the breath vs. heart rate graph. Choose the appropriate answer from the list and justify your choice.	Q2b: Which are the weather conditions under which it is possible to snow. Choose the appropriate answer from the list and justify your choice.	Q3b: Choosing the hardest rock from a list and explaining.	Q4b: Should we use mobile phones at school?

the levels as described in **Table 1** (also see **Appendix A** and **Appendix B** for questionnaires). In Question 1 the students were asked to choose the most convincing answer from the given list, and explain their choice for the most convincing arguments. In Question 2 the students were asked to choose the most convincing counter argument and explain their choice. The bold text in the brackets in the appendices corresponds to the level of argument based on the design of the assessment tool. Questions 3 and 4 were open-ended questions and the students were asked to provide, in writing, their argument or counter-argument. An overview of the questions in the tools is presented in **Table 2**.

Data Collection and Sample

Both tools were randomly assigned to students in each school. The questionnaires were administered to a total of 246 students (11-14 years old) in eight public and private schools in large city in the UK and suburbs, with 114 students completing Tool A and 134 students completing Tool B. The students came from different backgrounds, with most of them (86%) having English as their first language. The average age of students was 13, even though the ages ranged from 11-14, and most of the students were male (72% for Tool A and 67 for Tool B). This is since most of the private schools that participated in this study were boy schools. Based on teacher reports, none of the students in the sample were specifically taught how to write arguments, or how to evaluate arguments, and therefore the students were considered of equal ability in argumentation.

Data Analysis Process

The data analysis consisted of various stages, described in detail here:

Step 1: Each one of the questions was scored based on Toulmin's modified framework by Erduran and colleagues (Erduran et al., 2004) that is presented in the Theoretical Framework. For the first and second question (multiple choice questions) based on the design of the questions each one of the responses in the assessment tool corresponded to an argument level, and a score as presented in **Table 1** (also see supplementary materials for responses). The third and fourth questions were read and coded based on the categories in **Table 1**, and then scored using the same logic described above. Therefore, a response which included a claim, warrant, and one or more rebuttals, was considered as a Level 5 argument (the highest) and was given the highest score (4). On the contrary a response that included only a claim, or an irrelevant response was identified as Level 1 and was scored with 0. A representative example of a response from Test B, Question 4 that was coded as Level 4 (claim, warrant and data) and therefore received a score of 3 is the following:

'I believe that we should bring our mobile phones at school	claim
because if there is an emergency such as an unexpected after school activity	data
then you can phone whoever is picking you up to come at a later time'. (Student 98, Male)	warrant

Initially 20% of the responses on both assessment tools were coded independently by the first author and a second researcher with expertise in written argument. The initial analyses were discussed until agreement was reached. The first author then coded the remaining questionnaires and when all tests were scored various statistical analyses were then conducted to explore if the four items in each test correlated, to check the degree of reliability of the questions in each form. The inter-item correlation for both tests (Tool A and Tool B) were examined to look at their relationships. The results identified a low correlation that existed

Table 3. Cronbach's Alpha proposed modifications

	Cronbach's Alpha if question deleted for Tool A	Cronbach's Alpha if question deleted for Tool B
Q1	0.674	0.705
Q2	0.508	0.532
Q3	0.394	0.361
Q4	0.489	0.478

Table 4. Descriptive statistics of each question

Test	Score	Q 1		Q 2		Q 3		Q 4	
		N	%	N	%	N	%	N	%
A	0	0	0.000	8	6.957	14	12.174	21	18.261
	1	13	11.404	8	6.957	24	20.870	12	10.435
	2	37	32.456	52	45.217	56	48.696	57	49.565
	3	10	8.772	7	6.087	14	12.174	23	20.000
	4	54	47.368	40	34.783	7	6.087	2	1.739
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
		2.920	1.122	2.550	1.230	1.790	1.013	1.770	1.029
B	0	3	2.239	22	16.418	25	18.797	30	22.727
	1	0	0.000	20	14.925	8	6.015	3	2.273
	2	70	52.239	30	22.388	21	15.789	67	50.758
	3	18	13.433	34	25.373	34	25.564	26	19.697
	4	43	32.090	28	20.896	45	33.835	6	4.545
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
		2.731	0.990	2.194	1.368	2.496	1.480	1.811	1.133

between question 1 and the other questions of each form (**Table 3**). As a result, a decision was made to delete the specific question for both tests. After the deletion of Question 1, Cronbach's alpha was calculated, the degree of internal consistency was 0.674 for Test A and 0.705 for Test B, placing the degree of reliability of the scores from both tests in the acceptable range.

Step 2: In order to identify differences between the different sub-skills (i.e., content-specificity, differences between choosing an argument and writing an argument), the descriptive statistics of each item were initially examined (see **Table 4**) to identify possible differences in scores between the questions.

RESULTS

The purpose of this study was threefold: (a) to explore if there is a difference in the level of complexity in students written argumentation sub-skills (i.e., choosing a convincing argument, choosing a convincing counter-argument, writing an argument, writing a convincing counter-argument), (b) to explore if written argumentation is content specific, and (c) to explore the structure of students' written arguments and counterarguments.

Table 4 presents the descriptive statistics for the questions in both tests. The score for each question corresponds to levels of argument as presented in **Table 1** (i.e., the best argument received 5 points and the worst argument 0 points). For example, a question that was scored with 0 is a Level 1 question, and a question that was scored with 4 is a Level 5 question.

It is evident from the results in **Table 4** that choosing a convincing argument (Question 1) or a convincing counter-argument (Question 2) are easier sub-skills than writing an argument or a counter-argument. Specifically, the mean scores for questions 1 and 2 for both tests are higher compared to the mean scores for Questions 3 and 4 for Test A. Additionally for the first two questions for Test A there is a tendency for students to choose responses that are either Level 3 (data only), or Level 5 (data, warrant and rebuttal), but they do not choose responses that are at Level 4 (data and warrant). The same tendency was obvious for Question 1 in Test B, but not for Question 2 in which students choose Level 3, Level 4 and Level 5 responses in the same frequency.

Looking back at the results in **Table 4**, writing an argument or a counter-argument are more difficult sub-skills compared to choosing a convincing argument. Additionally, looking at the scores for Questions 3 and 4 for both tests it is evident that there is a tendency for students to write arguments that are Level 3 (data only,

Table 5. Inter-Item Correlation Matrix for Tests A (top triangle) and B (bottom triangle)

	Q1	Q2	Q3	Q4
Q1		.095	.187	.129
Q2	-0.002		.459	.317
Q3	.14	.483		.474
Q4	.126	.333	.524	

scored as 2), with the exception of Question 3, Test B which focused on a topic already taught to the students instead of an unknown topic. To further explore this issue, the correlation of the items in each questionnaire was calculated and is presented in **Table 5**.

As shown in **Table 5**, Question 1 does not correlate well with any of the other questions, in either of the two assessment tools. Furthermore, in both tools Question 2 correlated with Question 3, and Question 3 correlates with Question 4.

FINDINGS AND DISCUSSION

The first finding of this study suggests that there is a difference in the level of complexity in students' written argumentation sub-skills. More specifically, for four sub-skills that were evaluated in the tests – namely choosing a convincing argument, choosing a convincing counter-argument, writing an argument, writing a convincing counter-argument – the students had different degrees of success. Based on the results (**Tables 4** and **5**), the first question in both tests – the question designed to evaluate students' skill to choose a convincing argument – appears to be the easiest, and a different sub-skill. Furthermore, writing an argument or a counter argument are skills which appear to be more difficult than selecting the best argument from a given list (**Table 4**). The order of complexity for the sub-skills, starting from the easiest, is: choosing a convincing argument, choosing a convincing counter-argument, writing an argument, writing a convincing counter-argument. The degree of difficulty of writing counterarguments as opposed to writing arguments has been identified in previous studies with older students (Nussbaum & Schraw, 2010; Shehab & Nussbaum, 2014). Our contribution in this study is that we have identified the degree of complexity for all four sub-skills that were included in the test. By identifying that writing an argument is a more difficult skill to acquire, or that students are not acquainted with it, it can help educators to design better scaffolding structures to support students when writing counterarguments (Shehab & Nussbaum, 2014).

A second finding of this study associates with whether written argument is content specific. By comparing corresponding questions in the two tests (i.e., questions testing the same sub-skill but with different content) it is evident that the mean scores were different (see **Table 3**). Since the structure of the test, and the design were the same, the difference in the degree of difficulty for the two tests can be attributed to the difference in the content of the questions, or to students' familiarity with the content. Specifically, the first question in Test A was referring to the electric circuit that is taught in the curriculum, but in Test B the first question was a "breath vs. heart rate graph" which was not taught to the specific students, and also requires graph interpretation skills. The second question in Test A was about light and how it travels, and in Test B it was about weather conditions under which it can snow. The students were more familiar with the first topic since this was taught in the curriculum, and the mean was higher for that question. The third question in Test A was about mosquitos, a realistic problem, and in Test B the question was a graphic representation of an experiment with rocks, a topic that is taught in the curriculum. Therefore, the mean score appeared to be higher for the questions addressing topics that were either taught as part of the curriculum, or of personal interest. This finding is similar to previous studies, according to which students reason better when they have personal knowledge of the topic (i.e., Kuhn, 1991), or do not lack the scientific knowledge necessary to interpret the phenomenon (Van Aufschnaiter et al., 2008). A contribution arising from this finding is that this study evaluated written arguments both for scientific and everyday topics and concluded that success in writing about these types of topics is based on knowledge of the topic, regardless of whether this is a scientific or everyday topic.

A third finding of this study is associated with the quality of students' written arguments. Students tend to provide written arguments that are based either on warrants or data (Level 2 or 3 arguments – see Table 4, Question 2), without including rebuttals in them, a finding similar to previous studies (Kelly & Takao, 2002).

Research implications arising from the findings include exploring in detail how students choose to agree or disagree with given claims in different situations – for example exploring the difference in agreeing with media claims on socioscientific issues as opposed to scientific claims in the science classroom. Implications for teaching include using different teaching approaches for scientific and everyday argumentation.

CONCLUSIONS AND IMPLICATIONS

Given the emphasis in argumentation in science education in recent years, methodologically the assessment of argument has become one of the dominant issues in the field. However, assessment tools specially designed to evaluate students' written arguments have not been the emphasis of research studies. One of the main emphases of this study was designing a tool that can potentially evaluate students' ability to consider arguments, write them, and evaluate them as these are important scientific literacy skills in the 21st century. Overall, an additional goal was to explore possible differences between the sub-categories that were included in the argument assessment test – namely choosing a convincing argument, choosing a convincing counter argument, writing a convincing argument, and writing a convincing counter argument. Previous studies (Author, 2017) have shown that there is an uncertainty regarding the evaluation of arguments since teachers find it difficult to evaluate arguments in their classroom, mainly because of the complexity of the frameworks and the unavailability of assessment tools to assist them in this effort. Therefore, it is not easy for teachers, especially teachers that are in the early stages of incorporating argumentation to their teaching practices, to see the impact of their teaching on students' argument skills or even to evaluate them. Specifically, the lack of an argument assessment tool makes it difficult for the teachers to track the development of their students' argument skills over time, and therefore convince them of the impact of their teaching. A recommendation arising from this finding includes using the assessment tool with in-service teachers and their classes in order to identify whether in actual practice this tool can contribute to the evaluation of the students. Additionally, such a study will inform us of whether using this tool is practical from the perspective of the teachers, as well of whether this tool can track changes and development in students' argument skills, especially for those students that had high scores to begin with. Finally, a second recommendation involves using the tool in the context of teacher education, since research has shown that pre-service teachers have difficulties in not only understanding argumentation, but teaching argumentation in their classrooms (e.g., Martín-Gámez & Erduran, 2018). Such tools can support teachers in developing the necessary skills for teaching argumentation.

Author contributions: All authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. All authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Ethics declaration: Authors declared that ethical approval was not required since no personal information about the participants was collected. The questionnaires were sent to teachers via post and were returned to the researchers via post with no indication identifying the students or classroom.

Declaration of interest: Authors declare no competing interest.

Data availability: Data generated or analysed during this study are available from the authors on request.

REFERENCES

Allchin, D. (2022). Who speaks for science? *Science and Education, 31*, 1475-1492. https://doi.org/10.1007/s11191-021-00257-4

Allchin, D. (2023). Ten competencies for the science misinformation crisis. *Science Education, 107*, 261-274. https://doi.org/10.1002/sce.21746

Bell, P., & Linn, M. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education, 22*(8), 797-817. https://doi.org/10.1080/095006900412284

Bravo-Torija, B., & Jimenez-Aleixandre, M.P. (2018). Developing an initial learning progression for use of evidence in decision-making contexts, *International Journal of Science and Mathematics Education, 16*, 619-638. https://doi.org/10.1007/s10763-017-9803-9

Erduran, S. (2008). Methodological foundations in the study of argumentation in science classrooms. In S. Erduran & M. Jimenez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research*. Springer. http://doi.org/10.1007/978-1-4020-6670-2_3

Erduran, S. (2022). *Argumentation in chemistry education. Research, policy and practice*. Royal Society of Chemistry.

Erduran, S., Osborne, J., & Simon, S. (2004). TAPping into argumentation: Developments in the application of Toulmin. *Science Education, 88*(6), 915-933. https://doi.org/10.1002/sce.20012

Erduran, S., Ozdem, Y., & Park, J.-Y. (2015). Research trends on argumentation in science education: A journal content analysis from 1998–2014. *International Journal of STEM Education, 2*(1), 5. https://doi.org/10.1186/s40594-015-0020-1

Ferretti, R. P., & Graham, S. (2019). Argumentative writing: theory, assessment, and instruction. *Reading and Writing, 32*, 1345-1357. https://doi.org/10.1007/s11145-019-09950-x

Gleim, L. K., Sampson, V., Hester, M., Williams, K., Sanchez, J., & Button, E. (2010). *How middle school and high school students evaluate the claims and arguments found within articles written for the popular press: A comparison study* [Paper presentation]. International Conference of the National Association of Research in Science Teaching, Philadelphia, PA.

Jiménez-Aleixandre, M., & Pereiro-Munoz, C. (2002). Knowledge producers or knowledge consumers? Argumentation and decision making about environmental management. *International Journal of Science Education, 24*(11), 1171-1190. https://doi.org/10.1080/09500690210134857

Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, A. R. (2000). 'Doing the lesson' or 'Doing science': Argument in high school genetics. *Science Education, 84*(6), 757-792. https://doi.org/10.1002/1098-237X(200011)84:6%3C757::AID-SCE5%3E3.0.CO;2-F

Kelly, G., & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Science Education, 86*, 314-342. https://doi.org/10.1002/sce.10024

Kuhn, D. (1991). *The skills of argument*. Cambridge. https://doi.org/10.1017/CBO9780511571350

Lawson, A. (2003). The nature and development of hypothetico-predictive argumentation with implications for science teaching. *International Journal of Science Education, 25*(11), 1387-1408. https://doi.org/10.1080/0950069032000052117

Lederman, N. G. (1992). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching, 29*(4), 331-359. https://doi.org/10.1002/tea.3660290404

Lee, M., Wu, X., & Tsai, C. (2009). Research trends in science education from 2003 to 2007: A content analysis of publications in selected journals. *International Journal of Science Education, 31*(15), 1999-2020. https://doi.org/10.1080/09500690802314876

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

Martín-Gámez, C., & Erduran, S. (2018). Understanding argumentation about socio-scientific issues on energy: A quantitative study with primary pre-service teachers in Spain. *Research in Science and Technological Education, 36*(4), 463-483. https://doi.org/10.1080/02635143.2018.1427568

Mason, L., & Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. *Learning and Instruction, 16*(5), 492-509. https://doi.org/10.1016/j.learninstruc.2006.09.007

McNeill, K. L., & Krajcik, J. (2008). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching*, 45(1), 53-78. https://doi.org/10.1002/tea.20201

McNeill, K., & Pimentel, D. (2009). Scientific discourse in three urban classrooms: The role of the teacher in engaging high school students in argumentation. *Science Education, 94*, 203-229. https://doi.org/10.1002/sce.20364

Nussbaum, M., & Schraw, G. (2010). Promoting argument-counterargument integration in students' writing. *The Journal of Experimental Education, 76*(1), 59-92. https://doi.org/10.3200/JEXE.76.1.59-92

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*(2), 224-240. https://doi.org/10.1002/sce.10066

Norris, S. P., Phillips, L. M., & Korpan, C. A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest, and reading difficulty. *Public Understanding of Science, 12*(2), 123-145. https://doi.org/10.1177/09636625030122001

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S.-Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*(6), 821-846. https://doi.org/10.1002/tea.21316

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020. https://doi.org/10.1002/tea.20035

Papanastasiou, E. C. (2019). Participant pre-knowledge and attitudes in research. In J. E. Edlund & A. L. Nichols (Eds.), *Advanced research methods for the social sciences* (pp. 103-128). Cambridge University Press.

Perkins, D. N. (1993). Teaching for understanding. *American Educator, 17*, 28-35.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context bound? *Educational Researcher, 18*(1), 16-25. https://doi.org/10.3102/0013189X018001016

Rapanta, C., & Christodoulou, A. (2022). Walton's types of argumentation dialogues as classroom discourse sequences. *Learning, Culture and Social Interaction, 36*, 100352. https://doi.org/10.1016/j.lcsi.2019.100352

Rodríguez-Mora, F., Cebrián-Robles, D., & Blanco-López, A. (2022). As assessment using rubrics and the Rasch Model of 14/15-year-old students' difficulties in arguing about bottled water consumption. *Research in Science Education, 52*, 1075-1091. https://doi.org/10.1007/s11165-020-09985-z

Sampson, V., & Clark, D. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*(3), 447-472. https://doi.org/10.1002/sce.20276

Sampson, V., & Clark, D. (2011). A comparison of the collaborative scientific argumentation practices of two high and two low performing groups. *Research in Science Education, 1*(41), 63-97. https://doi.org/10.1007/s11165-009-9146-9

Sandoval, W. A., & Millwood, K. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23-55. https://doi.org/10.1207/s1532690xci2301_2

Sandoval, W. A., & Reiser, B. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education, 88*(3), 345-372. https://doi.org/10.1002/sce.10130

Sharon, A. J., & Baram-Tsabari, A. (2020). Can science literacy help individuals misinformation in everyday life? *Science Education, 104*, 873-894. https://doi.org/10.1002/sce.21581

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Van Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching, 45*(1), 101-131. https://doi.org/10.1002/tea.20213

van Eemeren, F. H., Grootendorst, R., Henkemans, F. S., Blair, J. A., Johnson, R. H., & Krabbe, E. C. W. (1996). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Mahwah, NJ, Lawrence Erlbaum Associates, Inc.

van Eemeren, F. H., & Grootendorst, R. (2003). *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616389

Voss, J. F., & Means, M. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction, 1*(4), 337-350. https://doi.org/10.1016/0959-4752(91)90013-X

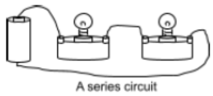
Zeidler, D. (1997). The central role of fallacious thinking in science education. *Science Education, 81*(4), 483-496. https://doi.org/10.1002/(SICI)1098-237X(199707)81:4<483::AID-SCE7>3.0.CO;2-8

APPENDIX A

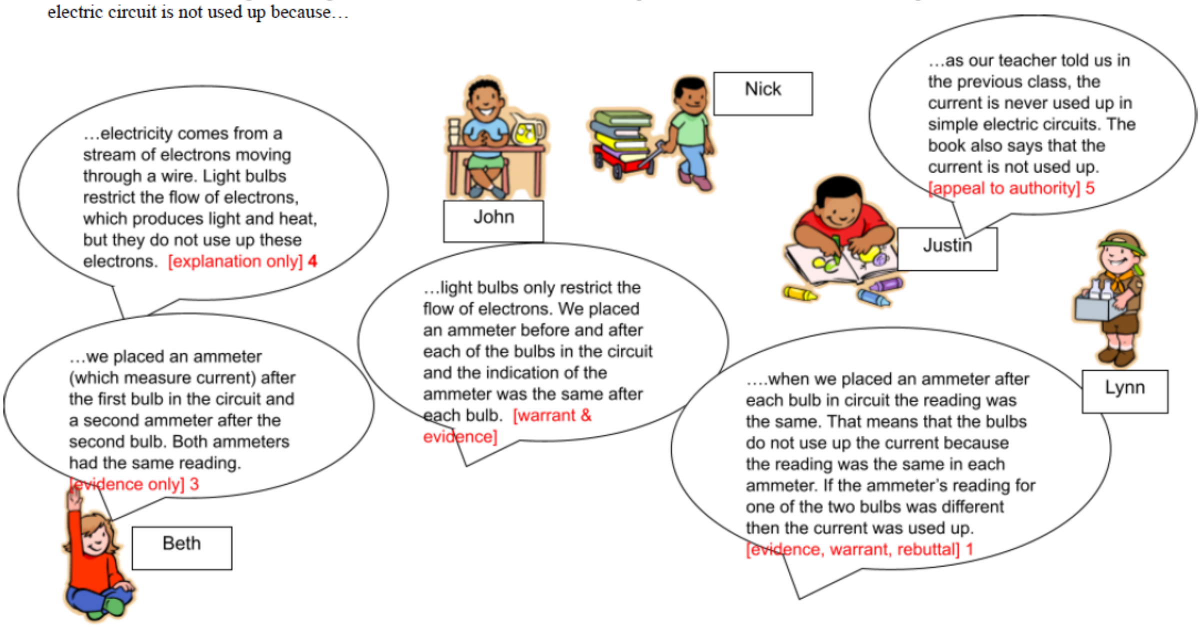
Question 1:

Directions: *In this question you are given a claim and 5 different ways that a person could justify the claim. You job is to rank each justification from 1 to 5: 1 being the most convincing justification and 5 being the least convincing justification.*

The figure at right illustrates a simple series circuit which consists of two light bulbs, some wire, and a battery.



Claim: The current that goes through the two identical bulbs shown in the picture (connected in series) in a simple electric circuit is not used up because…

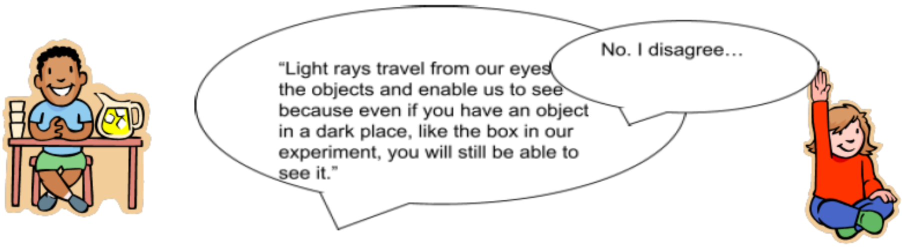


Then you have to explain your choice for the most convincing and least convincing justification.

Part 2:
I have chosen _______ to be the **most** convincing argument because…

Part 3:
I have chosen _______ to be the **least** convincing argument because…

Question 2:
Andrew and Kelly are discussing some of the experiments in their science class.



Your task is to rank these 6 different challenges in terms of how strong you think they are, hence what you rank as 1 is what you believe that Kelly should say. *Remember that you can only rank one challenge as 1, one challenge as 2, and so on.*

Part 1

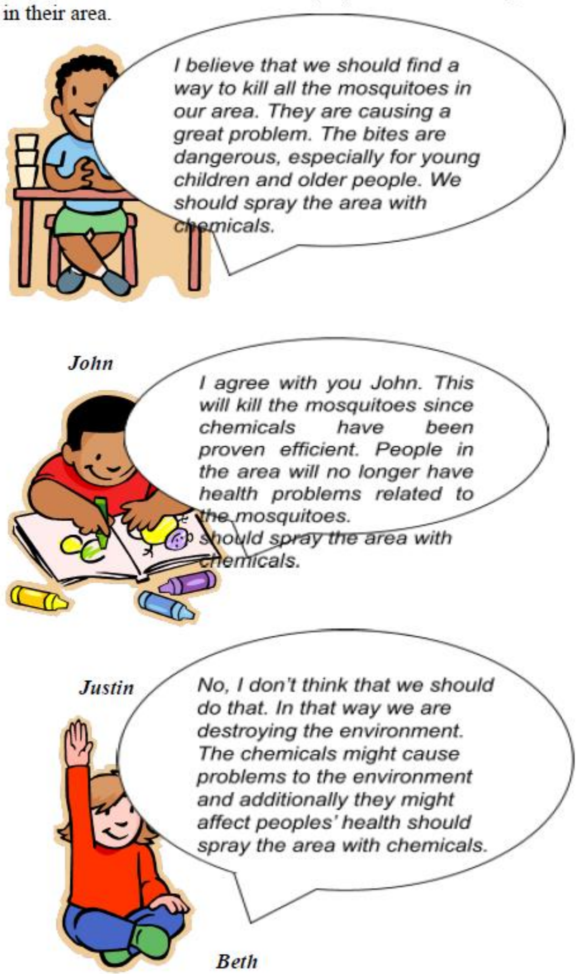
		Your ranking
(a)	…because the light does not travel from our eyes [counter-claim only]	5
(b)	…light does not travel from our eyes. Our eyes absorb light that is reflected off objects. [rebuttals against thesis no grounds]	4
(c)	…because your observation is wrong. When we were doing the experiment with the box we noticed that we could see the object in the box. We made a hole in the box and then we could see the object. This hole on the box was allowing the light could come in. [rebuttals against grounds no grounds]	2
(d)	…because you can not see an object in a completely darkness. When we had the object in the box and we were trying to see it from the small hole that was not possible. When we made another hole and the light could come in the box then we could see it. So it is not the light that travels from our eyes but the other way around. [rebuttals against grounds with grounds]	1
(e)	…the light does not travel from our eyes because if it did then we would be able to see all the objects. But when we had the object in the box and we were trying to see it from the small hole that was not possible because there was no light. [rebuttals against thesis with grounds]	3

Part 2:
I have chosen_______ to be the **most** convincing counter-argument because…

Part 3:
I have chosen_______ to be the **least** convincing counter-argument because…

Question 3:

John, Justin and Beth are discussing a problem concerning the increasing number of mosquitoes in their area.



I agree with ....because ....

....

....

....

....

....

....

Question 4:

The local council has decided to build a nuclear energy factory because a very small amount of nuclear fuel will provide a very large amount of electricity. The people in the area disagree. They prefer to use wind energy instead. You are representing the people in the area. Check the facts' cards below and try to write a counter-argument to convince the local council to use wind energy instead.

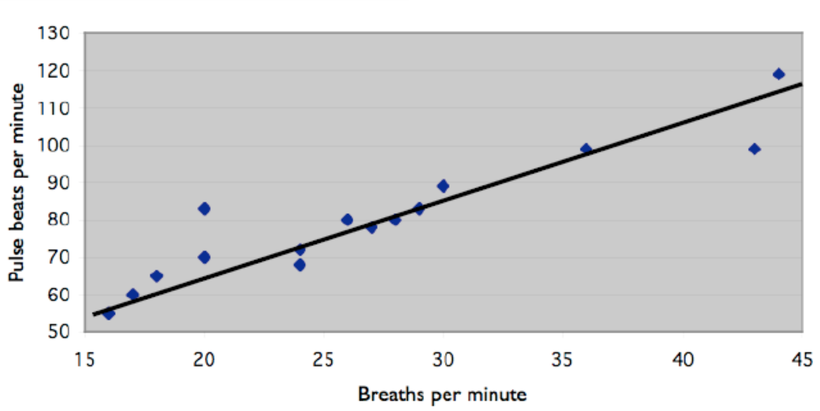
WIND ENERGY: FACTS	NUCLEAR ENERGY: FACTS
1. Relatively easy to set up.	1. Very expensive to set up.
2. Does not work when there is no wind or when the wind is too strong.	2. Can be very harmful if it goes wrong.
3. Each wind turbine does not generate very much electricity.	3. A very small amount of nuclear fuel (uranium) provides a very large amount of electricity.
4. Does not produce any gases that pollute the atmosphere.	4. Does not produce any gases that pollute the atmosphere.
5. Low maintenance costs - does not need any fuel.	5. Very high maintenance costs
6. Some people think wind turbines look ugly.	6. Very reliable and can provide energy for a long time

(example adapted by the IDEAS pack, Osborne et al., 2005)

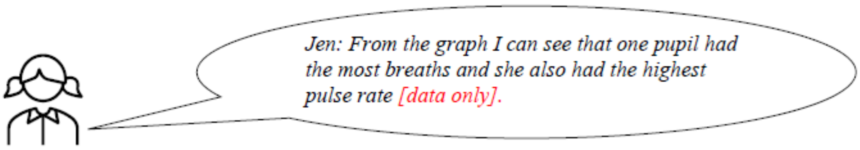
I believe that we should not use nuclear energy because ....
....
....
....
....
Instead, we should use wind energy because ....
....
....
....
....

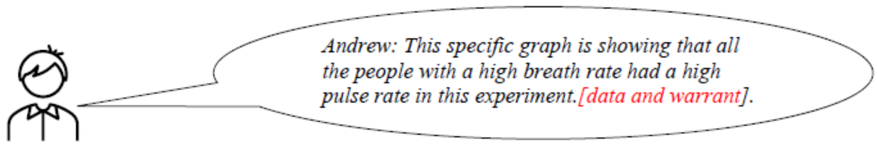
APPENDIX B

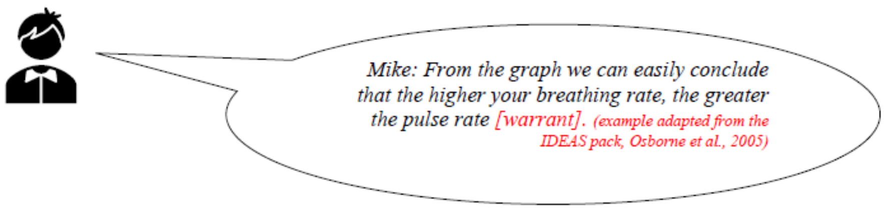
Question 1:
Some students were investigating whether there is a relationship between people's pulse rate and the number of breaths they take. The students recorded the pulse rate and breathing rate of 15 different people. The scatter graph for their results is shown below.

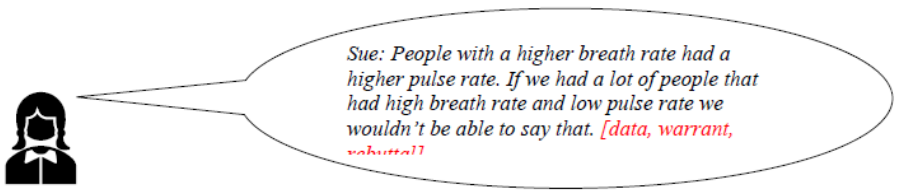


Jen, Andrew, Mike and Sue are trying to explain what this graph is showing.









Part 1:

I consider …………the most convincing explanation because ……………………………….

……………………………………………………………………………………………………......

……………………………………………………………………………………………………......

Part 2:

I consider …………the least convincing explanation because ……………………………….

……………………………………………………………………………………………………......

……………………………………………………………………………………………………......

Question 2:
The table below contains weather data from four different towns.

	Clouds in the sky	Highest Temperature
Town A	No	25 °C
Town B	Yes	30 °C
Town C	No	4 °C
Town D	Yes	-5 °C

Andrew: I think that it will definitely snow in Town C because there are clouds in the sky and because the temperature is low. I know that when the temperature is low, like 0°C or less, then it is always snowing.

His classmates disagree. Which of the following do you consider a convincing counter-argument?

Part 1	**Your ranking**
Dan: It will not snow in Town C but it is going to snow in Town D. It is also not possible to snow in Town A and Town B. But it is almost certain that it will snow in Town D because the temperature is below 0°C. We know that when the temperature is below 0°C and there is rain (from the clouds) then the rain will turn into snow. [rebuttals against thesis with grounds]	3
Helen: It will not snow in Town C. The temperature is not the only aspect that is related to snowing. I know that because if temperature was the only thing then we would have snow often in London since the temperature in the winter is often below 0°C. [rebuttals against grounds with grounds]	1
Mark: I am sure that you are wrong. It will snow in Town D for sure because the data show that. However it is unlikely to snow in Town C. Also I know that it is not probable to snow in Town A or Town B. [rebuttals against thesis no grounds]	4
Brad: Come on Andrew, you cannot be right. It does not always snow when the temperature is 0°C or less. Other variables influence as well. So you can not be that sure that it is going to snow in Town C. [counter-claim only]	5
Maria: Andrew I think that you are wrong, it will not snow in Town C because temperature is not the only factor for snowing. The probability of snowing in Town C is not that high at all. It is more likely to snow in Town D. [rebuttals against grounds no grounds]	2

Part 2:
I have chosen_______ to be the **most** convincing counter-argument because…

Part 3:
I have chosen_______ to be the **least** convincing counter-argument because…

Question 3: Jeremy tested four rocks to see how hard they are.

Jeremy rubbed each of them against some hard metal surface for a minute as shown below.

Here you can see pictures of how the rocks looked before and after the experiment.

	Before the test	**After the test**
Rock 1		
Rock 2		
Rock 3		
Rock 4		

Which is the hardest rock of the four?

…………………………………………………………………………………………………...

…………………………………………………………………………………………………...

I know that ………..is the hardest rock because ……………………………………………………

…………………………………………………………………………………………………...

…………………………………………………………………………………………………...

Additionally, I know that ……………………… can not be the hardest rock because …………...

…………………………………………………………………………………………………...

…………………………………………………………………………………………………...

Question 4:

The head teacher of the school is thinking of stopping you from bringing mobile phones to the school. Think of as many reasons as possible that would convince your headmaster to allow you to bring them.

I believe that we should bring our mobile phones at school because……………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………