A psychometric look on writing and evaluating arguments

Abstract

The purpose of this study was to design and validate an argument assessment tool that could easily be used by teachers to evaluate their students' argument skills in science. An additional purpose of the study was to explore the different characteristics of students' written arguments, for example different levels of complexity or sub-skills that might exist in written arguments, and the content dependency of arguments. In this paper we present two written argument tools that were designed for 11-14 year old students, the validation process, and the main outcomes from applying the assessment tools to 246 students in the UK. The analyses of the data from two versions of the questionnaire (Test A and Test B) show that the scores from both tests have a good degree of reliability when the first item is removed (.674 for Test A and .705 for Test B). Additionally 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 (choosing a convincing counterargument, writing an argument, writing a counter-argument). Finally, the results from the questionnaire support previous studies showing that argumentation is content specific, and that argument construction is easier when the students' have knowledge of the topic. Implications for research include the design of a new tool combining the questions from both tests, using aspects of the questionnaire to explore whether argumentation is context specific, and using the tool with teachers.

Keywords: argument, evaluation, written assessment, content-specific, scientific reasoning

Introduction

A major aim of science education (National Research Council, 2012) is for students to engage with scientific inquiry, and become scientifically literate. Scientific literacy involves amongst others being able to read and comprehend scientific articles and articles about science, and engage in discussions about the validity of the conclusions. Scientific argumentation, which makes 'the connection between claims and data through justifications, or the evaluation of knowledge claims in light of evidence, either empirical or theoretical' (Jiménez-Aleixandre & Erduran, 2008), is considered an important aspect of scientific inquiry, and should therefore be developed as part of science education. The recent National Research Council (2012) framework recognizes the importance of argument and argumentation and states that students 'should learn how to evaluate critically the scientific arguments of others and present counterarguments' (p.19), be able to write their own arguments and critically read media reports and evaluate them. Argument (product) and argumentation (process) have been prominent fields in the science education community for the past decades, and as a meta-analysis of research studies in the field of science education illustrates, most of the top ten highly cited papers from 2002 to 2007 were concerned with argument and argumentation (Lee, Wu, & Tsai, 2009). Studies with students in the past decades have ranged from the analysis of students' arguments and argumentation using a variety of methods and data sets – researchers have analysed interviews in which people engaged in argumentation (i.e. Kuhn, 1991), videos of students' constructing arguments during lessons (i.e. Erduran, Osborne, & Simon, 2004; Jiménez-Aleixandre & Pereiro-Munoz, 2002; McNeill & Pimentel, 2009), students' artifacts created during the instruction (Author, 2011; Bell & Linn, 2000), and written essays or texts (Jiménez-Aleixandre, Rodriguez, & Duschl, 2000; Kelly & Takao, 2002a; Kelly, Regev, & Prothero, 2008; Osborne, Erduran, & Simon, 2004; Sampson & Clark, 2010; Sampson, Grooms, & Walker, 2011). Summarizing, the aforementioned studies identify difficulties that students face with argument, and argumentation, and propose instructional approaches that can help enhance both written and oral argumentation in the classroom.

Even thought research in the field has informed us of students' difficulties with argument and argumentation, and of ways to support them to engage in the process of argumentation and constructing arguments, not much argumentation is happening in the science classes (Newton, Driver, & Osborne, 1999), for two main reasons: (a) teachers lack the training and resources necessary to support a change in their instructional practices, which will enable them to engage their students in the scientific practices of argument and argumentation (Dawson & Venville, 2010; McNeill & Pimentel, 2009; Zembal-Saul, 2009); and (b) teachers find it difficult to assess argument and argumentation since they are unsure of what should be assessed, and how to evaluate the change of students argument and argumentation over time (Author, 2011). Therefore, even though science education research has influenced policy (i.e. National Research Council, 2012), science education practice has yet to be reached. Hand, Yore, Jagger and Prain (2010), support this view by stating that there is a research-practice gap between the work of science education researchers and science teachers, and this gap according to Bulterman-Bos (2008) derives from the lack of compatibility between the purposes of science education research and the needs of science teachers. In order to address this gap, research 'should generate valid new understanding, provide clear indications to classroom teachers of how they might be able to improve their practice' (McIntyre, 2005), and present these new understandings in ways that make sense to the teacher. Based on this gap between research and practice in argumentation, our aim in this study was to design and validate an assessment tool that could easily be used by

teachers as a way to assess young students' (11-14 year old) written argument skills, and that could be used over time to track improvement. In our reading of the published studies in the field of science education, currently such a tool does not exist, a tool that can potentially be used as a measure of whether students improve or not their argumentation. An additional purpose of the study was to explore the different characteristics of students' written arguments, for example different levels of complexity or sub-skills that might exist in written arguments, and the content dependency of arguments. In the section that follows we attempt to unpack the theoretical framework of our work, and present an argument justifying the importance of this study.

Theoretical Framework

The theoretical framework for this study incorporates research of language in science education (writing and reading), argument and argumentation in science education, and argumentation as a reasoning skill. According to Norris and Phillips (2003), scientific literacy entails in its fundamental sense the capacity to construct and interpret science texts, and in its derived sense, the capacity to be knowledgeable about science topics, concepts, processes, and methods. Therefore written argument is part of scientific literacy. The production of written text played a central role in scientific communities (Bazerman, 1988), and 'science exists because scientists are writers and speakers, and as a shared form of knowledge, scientific understanding in inseparable from written and spoken work' (Montgomery, 1996). According to Author (2010), writing during science lessons is an important means of refining and coordinating new ideas with existing knowledge and constructing and evaluating new ideas (Rivard, 2004). Therefore, writing has three different aspects, (a) writing to learn (i.e. Yore, Bisanz, & Hand, 2003); (b) writing to reason (Wellington & Osborne, 2001), and (c) writing to communicate (Sutton, 1998). In this section we provide an overview of the theoretical framework pertaining to the aforementioned areas.

Argumentation and argument

Argumentation is part of the practice of science for evaluating, refining and establishing new theories (Duschl, 1990; HoltonBrush, 1996) and is therefore considered a core element of the scientific enterprise. One of the theoretical underpinnings for studying argumentation within the context of science education is that in school science students must develop an understanding of the scientific enterprise, such as the aims and purposes of scientific work (Driver, Leach, Millar, & Scott, 1996). 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), and is a specific form of talk that can enable students to communicate in the classroom in 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). In this study we take the stance of D.Kuhn (2008) who distinguishes between the written and oral, and defines *argumentation* as the oral process, and *argument* as the final, written product. Therefore, we define *argument* as the *written* set of claims, data, warrants, and backings that contribute to the content of an argument, and use (Toulmin, 1958)work as the main theoretical framework that guides our work.

In Toulmin's Argument Pattern (TAP), the essential elements are claims, data, warrants and backings. Toulmin defines data are '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; Jimenez-Aleixandre et al., 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 (Erduran et al., 2004; Erduran & Jiménez-Aleixandre, 2007; Jiménez-Aleixandre & Erduran, 2008; 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 (e.g. Erduran et al., 2004; Erduran 2008; Jimenez-Aleixandre et al., 2000), or the terms have to be better defined, using indicating words to identify when something is a claim, a warrant or a rebuttal. Duschl (2008) suggests that this characteristic has an adverse effect on interrater reliability and it should not be used in science education. However, Zohar and Dori (2003) and Erduran et al. (2004) in their effort to address this issue and increase the validity and reliability of Toulmin's framework have introduced *justifications* which are essentially a collapsed category for data, warrant and backings. 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 etc) and does not evaluate the content (Erduran, 2008; Sampson & Clark; 2008). So, even thought 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. In order 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 possesses 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, Simon and Osborne, 2004)

In designing and evaluating the written assessment 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. Jimenez-Aleixandre et al., 2000; Osborne et al., 2004), and it enables us to design an assessment tool that can help teachers evaluate their students' improvement of arguments easily.

Argumentation as reasoning and critical thinking

In our work we also view argumentation 'to be a fundamental tool of reasoning' (Voss & Means, 1991, p.337), and furthermore, we agree with Halpern (1997) who views argumentation as 'an inferential process by which a person, beginning with some given information or premise, makes an inference which enables that individual to reach a conclusion or provide some new (inferred) information that was not given' (Voss & Means, 1991, p.338). Therefore, arguing is part of informal reasoning (Mason & Scirica, 2006) and critical thinking. The complexity of defining critical thinking is associated with the fact that critical thinking has been informed both by studies in psychology and by studies in learning and teaching; areas in which critical thinking is defined in different ways (Pithers & Soden, 2000). On the one hand, studies in learning and teaching define critical thinking as the ability to identify questions and pursue knowledge by yourself, and to be able to present evidence to support one's arguments (Boekaerts, 1997; Gibbs, 1992). On the other hand, studies in psychology define critical thinking as helping students to engage in thinking for themselves (i.e. Perkins, 1993), without necessarily defining the exact way in which this outcome could occur. However, in spite of the different definitions, there is consensus about some of the characteristics identifying critical thinking and hence, critical thinkers. Critical thinking for example involves the application of principles of 'reasoning, fallacy detection or argument criticism, irrespective of the context in which they are exercised' (Winch, 2006, p.59). Furthermore, Winch argues that critical thinking skills are connected to the ability to form opinions and negotiate a way in the world and are, therefore, closely associated with the idea of autonomy as a central organizing feature of life in contemporary societies. Critical thinking is also connected to what Rath (1996) defines as good thinking: the ability to compare, observe, summarize and classify, suggest hypotheses, make decisions, create, gather and organize data or information and apply principles to new situations. According to Winch (2006), it is accurate to characterize critical thinking as a process that is mainly concerned with the identification, critique and construction of deductive argumentation. Summarizing, as derived from the definitions of critical thinking, critical thinkers are people who can construct and rebut arguments, and are skilful in identifying a problem and its associated assumptions; they can clarify and focus on the problem, and analyse, understand and make use of inferences, exercise inductive and deductive logic, as well as being able to judge the validity and reliability of the assumptions, sources of data or information available (Kennedy, Fisher, & Ennis, 1991). Therefore, the theoretical framework of our work has been influenced by the idea that argumentation is a reasoning skill, and part of the critical thinking process, that can be applied to different contexts, when the students have the basic content knowledge.

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 her study with people of different ages and three everyday scenarios 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 (Erduran et al., 2004; Jiménez-Aleixandre et al., 2000; Koswloski, 1996; Sandoval & Reiser, 2004). Van Aufschnaiter, Erduran, Osborne and Simon (2008), argue that "even scientists may not be able to engage in high level argumentation when confronted with an unfamiliar task. Hence in order 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, Means and Voss (1996) in a study with university students from an economics department showed that ability and knowledge did not yield any difference in students' reasoning. Additionally the same researchers describe a follow up case study 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). In spite of the contradicting findings in previous studies regarding the content specificity of reasoning skills, we assume a situated cognition stance in our work, 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.

Literature Review

Research in the construction of arguments was initiated in the early 1990s by Deanna Kuhn, and then was followed by a number of studies in science education, exploring students' difficulties both in argument and argumentation. Kuhn's outcomes highlight that: most people tend to be certain of their theories; even people who base their theories on pseudo-evidence believe that what they are saying is indeed genuine evidence; they tend to reason better on the subjects for which they have personal knowledge; they assimilate any new information in existing theories and they express considerable certainty that new evidence supports their theories; no gender differences were evident as far as the development of argumentative skills is concerned; people that have knowledge of the subject seem to be more able to provide an alternative theory. In summary, Kuhn's (1989a; 1991; 1993) findings imply that people display an epistemological naiveté in their argumentative skills. Furthermore, the argumentative skills of participants were "*elementary preconditions for rational argument"* and, therefore, were not fully developed. Finally, Kuhn's (1989b) results indicate that school made no difference to the development of argumentation skills after the end of junior high school. That last finding led Kuhn to state that development of argumentative skills could only be enhanced by the opportunity to engage in argumentative dialogue, as only this strategy "externalizes

argumentative reasoning and offers the exposure to contrasting ideas and the practice that may facilitate its development." Even though Kuhn's research is not directly associated with the current study – mainly because it is not placed in school settings – her work provides insight into difficulties that students have when constructing arguments and these difficulties are similar to the ones students have in the classroom when they engage in discussions (e.g. Bell, 2004; Driver et al., 2000; Jimenez-Aleixandre et al, 2000; Sadler, 2004; Sandoval, 2003). Kuhn's work has been critiqued by Koslowski (1996), whose research suggests that people have greater capabilities to use argumentation than Kuhn's work would suggest, nevertheless studies in science education have shown that students struggle with constructing arguments and evaluating claims (Driver et al, 2000) and often have difficulty justifying claims with evidence and evaluating arguments (Sadler, 2006). However, developing expertise requires a process of coordinating theory and evidence, which requires practice in the process of constructing reasoned arguments (Driver et al., 1996).

More specifically, 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. Kuhn, 1991; Pontecorvo, 1993), 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, 2010), and they focus on their claims without necessarily justifying them (i.e. Jiménez-Aleixandre et al., 2000). Sandoval (2003), in a study with high school biology students and written arguments, found that the students did not cite data in their written arguments, even though it was evident that they had used the data during their investigations. Furthermore, Sandoval and Millwood (2005) 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. Additionally, the same researchers found that the students reacted differently in terms 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 the conceptual understanding, but not the structure to the quality. Finally Kelly and colleagues (Kelly et al., 2008; Kelly & Takao, 2002b) also explored undergraduate students' written arguments during an oceanography lesson using a framework they designed based on Bazerman's (1988) work. 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.

Other than writing arguments, another aspect of scientific argumentation is evaluating written arguments. Studies in the area of evaluating written arguments are sparse, but studies from the field of scientific literacy are informative. Norris, Phillips and Korpan, (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 descriptions of methods as such. In a similar study

Norris and Phillips (2003b) 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, in a more recent study Gleim and his colleagues (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.

Purpose and Importance of Study

The purpose of this study was to design and validate an argument assessment tool that could easily be used by teachers to evaluate their students' argument skills in science. An additional purpose of the study was to explore the different characteristics of students' written arguments, for example different levels of complexity or sub-skills that might exist in written arguments, and the content dependency of arguments. More specifically there is a design and a research component in this study. The design component is associated with the design of the argument assessment tool, whilst the research questions guiding the second component are:

(1) Is the current assessment tool an appropriate tool for measuring students' argument skills in science*?*
(2) What can we say about 11-14 year old students' written argument skills based on this test (i.e. sub-skills, context-specificity)?

The importance of this study lies on the fact that there is a lack of an assessment tool that can easily assess young students' written arguments, and can easily be used by teachers in their everyday practices. Additionally, such a study can help us identify whether there are any sub-skills in terms of written arguments, and understand how to further support the students to develop them.

Methods

In order to test whether the current assessment tool is an appropriate tool for measuring students' argument skills, a mixed-methods approach was used. The initial pilot testing of the questionnaire was based on open-coding, whilst the validation of the assessment test was based on a statistical analysis of the responses provided. Several versions of the assessment tool were designed and pilot tested before finalizing the two versions presented in this paper (Test A and Test B). In this section we present and justify the structure of the two tests, the data collection process and data analysis.

Design of the assessment test

During the initial phase of the design of the questionnaire various questions were designed to assess four different aspect of argumentation, namely: (a) deciding what is a convincing argument; (b) deciding what is a convincing counter-argument; (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. 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. As argued in detail in previous sections, one of the reasons that students should develop their argument skills is to enable them to evaluate claims and data and decide, in their everyday life, whether an argument is valid or not (Millar & Osborne, 1998). The four different parts of the questionnaire are presented in Table 1, and for each one of the parts, the levels indicate the level of the argument based on a modified version the Erduran et al., (2004) framework to account for the fact that rebuttals are less often in written 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.

[INSERT TABLE 1 HERE]

It is important to note that the first two questions in both tests were multiple-choice questions. For these questions five specific answers were provided, with each answer representing one of the levels as described in Table 1. An example from Test A is presented in the supplementary material. In this question the students were asked to choose the most convincing answer from the given list. The red text in the brackets is the level of argument based on the design of the assessment test. Questions 3 and 4 were open-ended questions and the students were asked to write their argument or counter-argument. An example from Test B is presented in the supplementary materials. Table 2 below provides an overview of the parts of the two questionnaires, and the content for each of the questions.

[INSERT TABLE 2 HERE]

As described earlier, the two assessment tests were designed to measure the same construct – written argument. The original idea to include all 8 items in one assessment test was rejected during an initial data collection phase. Initially, a version of the assessment test, consisting of 8 questions, two for each sub-skill, was designed and 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 (Test A and Test B).

Data Collection and Sample

One of the assumptions in this study is that both tests measure the same construct, written argument. The questionnaires were administered to a total of 246 students (11-14 years old) in eight public and private schools in London and suburbs, with 114 students completing Test A and 134 students completing Test B. The students came from different backgrounds, with most of them (86%) having English as their first language. Additionally, those students that came from

private schools had better overall performance in science, language and mathematics, even thought individual scores were not available for all students to compare. The average age of the sample was 13, even thought the ages ranged from 11-14, and most of the students were male. This is due to the fact that most of the private schools that participated in this study were boy schools. None of the students in the sample were specifically taught how to write arguments, or how to evaluate arguments. Table 3 presents descriptive information for the sample.

[INSERT TABLE 3]

Data Analysis Process

The data analysis consisted of various stages in order to identify: (1) whether the assessment test is a credible and reliable tool for measuring written argument, and (2) whether there are any differences in students written argument sub-skills. The various data analysis steps are described here in detail:

Step 1: In order to assure that the assessment tests were credible a number of measures were taken. Credibility of an instrument (Trochim & Donnelly, 2006), is its ability to measure the constructed reality of the participants and is measured in the following ways:

- Content validity: This type of validity measures if the constructs in the assessment test were well defined and based on a theoretical framework. The constructs in the current assessment test are associated with argumentation (claim, data, warrant, rebuttal), as defined by Toulmin's framework, presented in the theoretical framework of the paper. It is argued that the constructs are based on a sound theoretical framework and hence the assessment test is valid in terms of content validity.
- Face validity: This type of validity measures if the participants ascribe the same meaning to the items as the researchers do. For the purposes of the current study, 4.9% of the population (12 students) were interviewed to see if they understand the assessment test as intended. Based on the interviews, the responses provided both for the questionnaire and the interview were the same.
- Translational validity: This type of validity measures if the constructs are accurately translated into operational items based on expert opinion. For this part two experts in argumentation read the assessment tests to see if the items are written in a way that accurately reflects the theoretical constructs.

Step 2: In order to identify whether the two versions of the written argument test were valid, each one of the questionnaires was scored based on Toulmin's framework. For the first and second question the process was straightforward since each one of the responses in the assessment test corresponded to an argument level, and a score as presented in Table 1. 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 because if there is an emergency such as an unexpected after school activity then you can phone who ever is picking you up to come at a later time. Also if, for example, you get locked in the toilet, then you can contact the school and they would unlock the toilet for you.' (Student 98, Male, Private School)

Initially 20% of both assessment tests were coded independently by the first author and a second researcher with expertise in argument. The initial analysis was discussed until agreement was reached. The first author then coded the remaining questionnaires independently. When all tests were scored, various statistical analyses were conducted to explore if the four items in each test correlated, and to check the reliability of the questions and tests.

Results and Findings

A breakdown of the various scores obtained for the four items on each of the tests is presented in Table 4, in addition to the mean and standard deviation of each item. In Test A, item 1 was the easiest item with an average score of 2.920 (sd=1.112), while the most difficult item was item 4 with an average score of 1.770 (sd=1.029). In Test B, the pattern was similar with item 1 being the easiest item (x=2.731, sd=0.990) and item 4 being the most difficult one (x=1.811, sd=1.133).

[INSERT TABLE 4]

In order to identify whether the two versions of the assessment tool are appropriate for measuring students' argument skills Cronbach's alpha was calculated. Cronbach's alpha is a measure of the internal consistency of the items, and for Test A was .596, and for Test B was .616. As evident by Cronbach's alpha for both tests, the degree of internal consistency was relatively low in the two tests, with Test B having a slightly higher internal consistency than Test A. Therefore, the inter-item correlation for both tests (Test A and Test B) were examined in an attempt to identify a probable source for the low degree of internal consistency. The inter-item correlations for Test A that are presented in Table 7 reveal that the correlations range from 0.095 to 0.474. Although the correlations between items 2, 3, and 4 were all in the acceptable range (0.317-0.474), the correlations of item 1 with the other items were very low. An examination of the internal consistency if the specific item was deleted, further verified the inappropriateness of item 1 in Test A, since this statistic indicated that the reliability of the test would increase to 0.674 if item 1 were deleted (Table 8). Therefore, a decision was made to delete item 1 from Test A, leaving the test with 3 items. The inter-item correlations for Test B (Table 8) revealed a similar pattern of results as Test A. More specifically, although the intercorrelations between items 2-4 were in the acceptable range, the correlations of item 1 with the other items were very low, ranging from -0.002 to 0.140. A further examination of Cronbach's alpha if item is deleted statistic also showed that the internal consistency of Test B would increase if item 1 were deleted.

[INSERT TABLE 5]

[INSERT TABLE 6]
[INSERT TABLE 7]

[INSERT TABLE 8]

Once the scores from Item 1 were removed from both tests, the Cronbach's alpha was calculated again as 0.674 for Test A and 0.705 for Test B. This placed the degree of reliability of the scores from Test A and Test B in the good range. Therefore, summarizing the findings for the first research question, based on the analyses, the two versions of the current assessment tool are appropriate for measuring written argument in science if the first item from each test is removed. Specifically, if item 1 is removed then both tests are reliable with a Cronbach's alpha in the good range – .674 for Test A and .705 for Test B.

Summarizing the findings for the second research question: (a) the scores in the tests differed depending on the type of question – the 'write an argument' questions (Q3 and Q4) in both tests had lower mean scores, while the 'choosing an argument' (Q1 and Q2) had higher mean scores, (b) the scores differed based on whether the students were familiar with the topic or not, with familiarity (either knowledge from school curriculum or personal interest) being a predictor of a higher mean score; and (c) even thought the mean scores for Q1 and Q2 in both tests are similar, Q1 does not correlate well with any of the other questions, while Q2 does. Therefore, choosing a convincing argument appears to be a different skill than any other of the skills in the test; and (d) students tend to provide arguments that include either warrants only, or data only for the 'write an argument and counter-argument' questions (see Table 4).

Discussion and Implications

Given the emphasis in argumentation in science education in the 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 to our knowledge. One of the main emphases of this study was on designing a tool that can potentially evaluate students' ability to consider arguments, write them, and evaluate them. 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. Our first research question focused on examining whether the two versions of the assessment tool that were designed are reliable for measuring written argument for 11-14 year old students. Summarizing the findings, both versions of the tests are reliable if the first item from each test is removed. Therefore, the first contribution of this study is that two assessment tests that can be used researchers and teachers to evaluate written argument were designed and validated. Previous studies (Author, 2011) 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. 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. 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 research implication arising from this finding includes using the assessment tool with in-service teacher 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.

The second aim of this study was to explore the characteristics of students' written argument skills based on the findings from the tests. From the analyses of the two questionnaires it was evident that the first item in both tests was measuring something different from all the other items, since item 1 in both tests did not correlate well with any of the other items (see Tables 5-7). Therefore, choosing a convincing argument from a given list can be seen as a different type of skill than choosing a convincing counter-argument, writing an argument or writing a counter-argument. This is a contribution of the current study since previous studies did not identify any differences between these sub-skills of argument. Furthermore, the first question in both tests appears to be the easiest with a mean score of 2.920 for Test A, and 2.731 for Test B (see Table 4). Comparing the mean scores for each question for Test A and Test B, it is evident that in Test A the first and second questions were the easiest, while the third and fourth questions, the open-ended questions were more difficult. Therefore, writing an argument or a counter argument are skills which appear to be more difficult than selecting the best argument from a given list. To our reading of the literature, especially in the area of science education, this finding has not been listed in any previous studies, and therefore this is the second contribution of our study. By identifying that writing an argument is a more difficult skill to acquire, or that students are not acquainted with it, it can help us as educators and policy makes to place an emphasis on promoting this skill in the science classroom.

A third finding of this study is associates with the domain knowledge of the questions. Comparing corresponding questions in each one of the tests it is evident that the mean scores were different. Since the structure of the questionnaire, 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 is not a known topic, 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 presenting an experiment with rocks, a topic that is taught in the curriculum. The mean scored appeared to be higher for the questions addressing topics that were either taught, 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. A fourth finding of this study is that students tend to provide written arguments that are based either of warrants or data, without including rebuttals in them, a finding similar to previous studies (Kelly et al., 2008; Kelly & Takao, 2002b). Research implications arising from the findings of the second reason question 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.

Concluding, a good reasoner according to Means and Voss (1991) should be able to 'consider arguments counter to his or her argument and be able to refute them or to re-evaluate one's own position in reference to them […], a good reasoner should be able to evaluate arguments' (p.341). Based on the findings of the current study, and supported by previous studies, constructing arguments or evaluating them are not skills that are necessarily already developed for young students, but knowledge of the topic can help improve the quality of the arguments. Therefore, in order to improve the skills of argument, the knowledge of science should also be improved.

References

Author (2010)
Author (2011)
Bazerman, C. (1988). Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science. (01 October 1988).
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.
Boekaerts, M. (1997). Self-regulated learning: a new concept embraced by researchers, policy makers, educators, teachers, and students. Learning and Instruction, 7(161-186).
Bulterman-Bos, J. (2008). Will a clinical approach make education research more relevant for practice? Educational Researcher.
Dawson, V., & Venville, G. (2010). Teaching strategies for developing students' argumentation skills about socioscientific issues in high school genetics. Research in Science Education, 40(2), 133–148.
Driver, R., Leach, J., Millar, R., & Scott, P. (1996). Young people's images of science. Open University Press.
Duschl, R. (1990). Restructuring science education. Teachers College Press.
Duschl, R. (2008). Quality Argumentation and Epistemic Criteria. In S. Erduran & M. Jiménez-Aleixandre (Eds.), Argumentation in Science Education: Perspectives from Classroom-Based Research. Springer.
Erduran, S., & Jiménez-Aleixandre, M. (2007). Argumentation in Science Education. Springer Verlag.
Erduran, S., Osborne, J., & Simon, S. (2004). TAPping into Argumentation: Developments in the Application of Toulmin. Science Education, 88(6), 915–933.
Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. Science Education, 92(3), 404–423. doi:10.1002/sce.20263
Gibbs, G. (1992). Improving the Quality of Students' Learning. Bristol: Technical and Educational Services.
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 study1.
Halpern, D. F. (1997). Critical thinking across the curriculum. Lawrence Erlbaum.
Hand, B., Yore, L., Jagger, S., & Prain, V. (2010). Connecting research in science literacy and classroom practice: a review of science teaching journals in Australia, the UK and the United States, 1998–2008. Studies in Science Education, 46(1), 45–68. doi:10.1080/03057260903562342

Holton, G., Brush. (1996). Physics, the Human Adventure: From Copernicus to Einstein and Beyond. New Jersey: Rutgers University Press.
Jiménez-Aleixandre, M., & Erduran, S. (2008). Argumentation in Science Education: An overview. In S. Erduran & M. Jiménez-Aleixandre (Eds.), Argumentation in Science Education: Perspectives from Classroom-Based Research (pp. 3–27). Springer.
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.
Jiménez-Aleixandre, M., Rodriguez, M. P., & Duschl, R. (2000). "Doing the Lesson" or 'Doing Science': Argument in High School Genetics. Science Education, 84(6), 757–792.
Kelly, G., & Takao, A. (2002a). Epistemic Levels in Argument: An Analysis of University Oceanography Students' Use of Evidence in Writing. Science Education, 86, 314–342.
Kelly, G., & Takao, A. (2002b). Epistemic Levels in Argument: An Analysis of University Oceanography Students' Use of Evidence in Writing. Science Education, 86, 314–342.
Kelly, G., Regev, J., & Prothero, W. (2008). Analysis of lines of reasoning in written argumentation. In S. Erduran & M. Jiménez-Aleixandre (Eds.), Argumentation in Science Education (pp. 137–157). Argumentation in science education.
Kennedy, M., Fisher, M. B., & Ennis, R. H. (1991). Critical Thinking: Literature Review and Needed Research. In L. Idols & F. Jones (Eds.), Educational Values and Cognitive Instruction: Implications for Reform. Educational values and ….
Koswloski, B. (1996). Theory and Evidence: The Development of Scientific reasoning. Massachusetts: MIT Press.
Kuhn, D. (1989a). Children and Adults as Intuitive Scientists. Psychological Review, 96(4), 674–689.
Kuhn, D. (1989b). Children and Adults as Intuitive Scientists. Psychological Review, 96(4), 674–689.
Kuhn, D. (1991). The Skills of Argument. Cambridge.
Kuhn, D. (1993). Science as Argument: Implications for Teaching and Learning Scientific Thinking. Science Education, 77(3), 319–337.
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. doi: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. doi: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.
Lehrer, R., & Schauble, L. (2006). Cultivating model-based reasoning in science education. Cambridge handbook of the learning sciences.
Mason, L., & Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. Learning and Instruction, 16(5), 492–509.
McIntyre, D. (2005). Bridging the gap between research and practice. Cambridge Journal of Education, 35(3), 357–382. doi:10.1080/03057640500319065
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.
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.
Means, M. L., & Voss, J. F. (1996). Who reasons Well? Two Studies of Informal Reasoning Among Children of Different Grade, Ability, and Knowledge Levels. Cognition and Instruction, 14(2), 139–178.
Millar, R., & Osborne, J. (1998). Beyond 2000 Science education for the future; a report with ten recommendations - OpenGrey. King's College London.
Montgomery, S. L. (1996). Montgomery (1996) The scientific voice.
Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. International Journal of Science Education, 21(5), 553–576.
Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. Science Education, 87(2), 224–240. doi: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. doi:10.1177/09636625030122001
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. doi:10.1002/tea.20035
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.
Pithers, R. T., & Soden, R. (2000). Critical thinking in science education: a review. Educational Research, 42(3), 237–249.
Pontecorvo, C. (1993). Social interaction in the acquisition of knowledge. Educational Psychology Review, 5(3), 293–310.
Rivard, L. (2004). Are Language Based Activities in Science Effective for All Students, Including Low Achievers? Science Education, 88, 420–442.
Sadler, T. (2006). Promoting Discourse and Argumentation in Science Teacher Education. Journal of Science Teacher Education, 17(323-346), 323–346.
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. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/sce.20276/abstract
Sampson, V., & Clark, D. (2010). A Comparison of the Collaborative Scientific Argumentation Practices of Two High and Two Low Performing Groups. Research in Science Education, 1(41), 63–97. Retrieved from http://www.springerlink.com/content/c28324w66g588858/
Sampson, V., Grooms, J., & Walker, J. (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.
Sandoval, W. A., & Millwood, K. (2005). The quality of students' use of evidence in written scientific explanations. Cognition and Instruction, 23(1), 23–55.
Sandoval, W. A., & Reiser, B. (2004). Explanation-Driven Inquiry: Integrating Conceptual and Epistemic Scaffolds for Scientific Inquiry. Science Education, 88(3), 345–372.
Standards, C. O. C. F. F. T. N. K.-1. S. E., National Research Council. (2012). A Framework for K-12 Science Education. National Academies Press.

Sutton, C. (1998). New Perspectives on Language Learning. In B. Fraser & K. Tobin (Eds.), The International Handbook of Science Education (pp. 27–38). London: Kluwer.

Toulmin, S. (1958). The uses of argument. Cambridge: Cambridge University Press. Retrieved from http://books.google.com/books?id=LO25ZwEACAAJ&dq=inauthor:Toulmin+(1958+intitle:the+uses+of+argument)&cd=2&source=gbs_api

Trochim, W. M. K., & Donnelly, J. P. (2006). Research Methods Knowledge Base. Atomic Dog Pub Incorporated.

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.

van Eemeren, F. H., & Grootendorst, R. (2003). A Systematic Theory of Argumentation: The pragma-dialectical approach - Frans H. van Eemeren, Rob Grootendorst - Google Books. Cambridge University Press.

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.

Voss, J. F., & Means, M. (1991). Learning to reason via instruction in argumentation. Learning and Instruction, 1(4), 337–350.

Wellington, J. J., & Osborne, J. (2001). Language and literacy in science education. Open Univ Pr.

Winch, C. (2006). Education, Autonomy and Critical Thinking. London: Routledge.

Yore, L., Bisanz, G., & Hand, B. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. International Journal of Science Education, 25(6), 689–725.

Zeidler, D. (1997). The central role of fallacious thinking in science education. Science Education, 81(4), 483–496.

Zembal-Saul, C. (2009). Learning to Teach Elementary School Science as Argument. Science Education, 93(687-719).

Zohar, A., & Dori, Y. J. (2003). Higher Order Thinking Skills and Low-Achieving Students: Are They Mutually Exclusive? Journal of the Learning Sciences, 12(2), 145–181. doi:10.1207/S15327809JLS1202_1