Chemistry Education Research and Practice

PAPER

View Article Online
View Journal | View Issue

Cite this: *Chem. Educ. Res. Pract.*, 2017, **18**, 127

Research on evaluation of Chinese students' competence in written scientific argumentation in the context of chemistry

Yang Deng* and Houxiong Wang*

Attending to practice has become a significant topic in science education today. As scientific argumentation is a typical form of scientific practice as well as an important educational practice, more and more attention has been paid to it by science education researchers. Evaluating students' competence in scientific argumentation is one of the most important research topics, but in China, science researchers seldom concentrate on it because the diverse educational values of scientific argumentation need to be further understood. The present study sought to examine the performance of Chinese students participating in written scientific argumentation in the context of chemistry. After clarifying the conception of scientific argumentation in science education, and comparing the evaluation criteria in domestic and international science education research, written scientific argumentation tasks in the context of chemistry were designed and criteria for their evaluation were constructed and improved. In total five tasks were designed for evaluation. All of the five tasks were aimed at evaluating students' competence of selecting (or putting forward) claims, evidence and warrants, in addition, two tasks investigated the competence of refuting arguments. The general criteria for evaluation was constructed according to the four dimensions of scientific argumentation, they were the structure components, the content quality, the logic of justification and language. For each task, content criteria and performance criteria for evaluation were constructed. After analysis and improvement of the criteria based on two pilot tests and the Rasch model, it was obvious that the criteria met the standards, effectively and credibly, for this study on the assessment of students' competence in written scientific argumentation. The number of students who participated in the formal test was 578 (304 males and 274 females). Through this kind of evaluation, this study found that the students' competence in written scientific argumentation was generally weak, and was influenced by some factors. Specifically, firstly, the students could put forward claims and evidence more easily than warrants and rebuttals. Secondly, the specific tasks had an influence on the performance of the students in written scientific argumentation. In regard to other factors, gender did not influence the students' competence in written scientific argumentation, but the grade level and school level were key factors. The students' competence in written scientific argumentation at grade level four and three other school grade levels were significantly different. Finally, some changes to the Chinese chemistry curriculum were proposed based on the results of this study.

Received 29th March 2016,
Accepted 1st November 2016

DOI: 10.1039/c6rp00076b

www.rsc.org/cerp

Introduction

Attending to practice has become a significant topic in science education today. As scientific argumentation is a typical form of scientific practice as well as an important educational practice (Driver *et al.*, 2000; Erduran *et al.*, 2004; Erduran, 2007), and several new instructional methods and curricula have been developed over the last decade to help students to acquire the competence needed to participate in scientific argumentation (Zembal-Saul, 2009; Berland and Reiser, 2011; Walker *et al.*, 2011; Cavagnetto and Hand, 2012; Foong and Daniel, 2012; Hong *et al.*, 2013), more and more attention has been paid by science education researchers on evaluating students' competence in scientific argumentation (Kelly and Takao, 2002; Sandoval and Millwood, 2005; Ryu and Sandoval, 2012; Sampson *et al.*, 2012; Wu and Tsai, 2012; Mendonca and Justi, 2014). These types of studies can help us to examine how these new instructional methods and curricula work. Written argumentation is a form of argumentation which needs written language, it plays an important role in science because science can be treated as a form of language expression, especially the

Department of Chemistry, Central China Normal University, Wuhan, Hubei 430079, China. E-mail: yangdeng@mail.ccnu.edu.cn, 67127513@163.com

written language. As noted by Yore *et al.* (2006), written communication provides a means of articulating evidence, warrants, and claims; reflecting on proposed ideas; critiquing the scientific work of others; and establishing proprietorship of intellectual property.

In China, science researchers seldom concentrate on evaluating students' competence of scientific argumentation, because the diversified educational values of scientific argumentation need to be further understood. The present study sought to examine the performance of Chinese students participating in written scientific argumentation in the context of chemistry, it is useful to understand what kinds of barriers they may meet in scientific argumentation, and may give some suggestions for designing instructional methods and curricula specific to Chinese students.

Theoretical framework

Scientific argumentation

Argumentation is a human practice that we are all familiar with. Researchers in informal logic have defined the word "argumentation" in different ways, for example, Browne and Keeley (1998) mentioned that argumentation is equal to reasons plus conclusions, it is committed to produce reasonable links between reasons and conclusions. Some other researchers have noted that argumentation is a kind of social practice, which is a communicative and interactive act aimed at resolving a difference of opinion with the addressee by putting forward a constellation of propositions the arguer can be held accountable for to make the standpoint at issue acceptable to a rational judge who judges reasonably (van Eemeren *et al.*, 2014). Jiménez-Aleixandre and Erduran (2007) argued that argumentation can be defined in terms of both an individual or structural way as well as a social or dialogic way. The dialogic or social perspective on argumentation focuses on the interactions between two or more individuals in which the participants try to persuade or convince each other of the validity of their claims (McNeill, 2011).

Argumentation also plays an important role in science. In the perspective of positivism, science is not only a process of confirmation which needs evidence from experience and reasons from theories, but also a way of criticism, as illustrated by Popper's formula "P1 → TT → EE → P2···", which shows that science cannot develop without refuting, and is also part of argumentation. Many researchers who have devoted themselves to the sociology of science have claimed that science cannot be done without social construction. In terms of the social constructivism of science, scientific communities are important, new scientific knowledge is based on the common knowledge of the scientific community, and the conversation between different communities also influences scientists. Mercer (2000) identified three different types of discourse, they are disputational, cumulative and exploratory talk. Disputational talk is competitive, the goal is to stress the difference in opinion rather than to provide solutions to the argument. Cumulative talk is agreement in nature, the typical features of cumulative talk are repetition, confirmation and elaboration. Exploratory talk involves the presentation of points of view backed up by arguments and critical but constructive discussions about the different ideas. Discourse analysis is helpful in scientific argumentation analysis, because science itself is a kind of language expression. Warren and her colleagues (Warren *et al.*, 2001) argued for a "view [of] scientific meaning making as encompassing a varied complex of resources, including practices of argumentation and embodied imagining, the generative power of everyday experience, and the role of informal language in meaning making" (p. 532). For analysis of the discourse of science, the orientation of rhetoric was welcome in modern philosophy of science, because not only the expression of science, but also the discovery and the controversy of science are all rhetorical (Li, 2004). When scientists adopt the methods of rhetoric, it is useful to strengthen the reasonability of science. In a word, the actual scientific argumentation could not occur before a scientist acquired and applied scientific language.

Scientific argumentation in science education

Scientific argumentation is also a typical form of scientific practice as well as an important form of educational practice reflecting multiple theoretical domains. As Jiménez-Aleixandre and Erduran (2007) mentioned, when students participate in scientific argumentation, they can acquire access to the cognitive and metacognitive processes characterizing expert performance and enabling modeling, they can develop communicative competences, particularly critical thinking, achieve scientific literacy and the competences of talking and writing science, they can become enculturated to the practices of the scientific culture and the epistemic criteria for knowledge evaluation, and they can develop reasoning, particularly the choice of theories or positions based on rational criteria.

When arguing in science, students should treat themselves as knowledge constructors, collaborative competitors and critical reflectors. But many science researchers have found that students may also meet some difficulties in scientific argumentation. Generally speaking, students may experience four kinds of difficulties. The first difficulty is recognizing the means or the goals of argumentation. Osborne *et al.* (2012) argued that "as students have no sense of how scientific knowledge is constructed, they do not see the value that critique and argumentation might have in establishing a secure basis for belief" (p. 11). Sampson *et al.* (2011) suggested that "students' not understanding the goals and norms of scientific argumentation and how these goals and norms diverge from the forms of argumentation they are accustomed to rather than a lack of skill or mental capacity" (p. 224). The second difficulty is understanding and using evidence to support a claim. As Garcia-Mila and Andersen (2007) mentioned, students may have strong confirmation biases either selecting evidence to confirm prior theories or assessing them differently (or even ignoring prior theories), according to whether the evidence confirms or disputes prior theories, jumping to conclusions before enough evidence is available. Other researchers have also reported that students were not able to choose appropriate evidence to support their claims, some evidence was irrelevant, some was not sufficient

(McNeill and Krajcik, 2007), and sometimes they could not even distinguish between evidence and claims (Kuhn, 1989). The next difficulty students may meet in scientific argumentation is reasoning. They may not connect evidence and claims correctly, sometimes they only use logical reasoning (Kelly and Takao, 2002), even experience (Erduran *et al.*, 2004). The fourth difficulty is rebuttal. When they participate in scientific argumentation, students seldom take rhetoric into consideration, they do not recognize the value and function of persuasion (Sandoval and Millwood, 2007) and only emphasis their own claims (Pontecorvo and Girardet, 1993).

Written scientific argumentation

It is possible to classify scientific argumentation into written scientific argumentation and oral scientific argumentation based on the criterion of the form of expression. Written scientific argumentation uses written language to present all of the components and the process of argumentation, a research paper published in a journal, book, or the Internet can all be treated as written scientific argumentation completed by scientists. When taking part in written scientific argumentation, a person must be equipped not only with scientific knowledge, but also the knowledge and the ability of writing, such as the knowledge of vocabulary, grammar and rhetoric, the ability to read, organize passages, evaluate and criticize others. Sometimes, they should also understand and apply specific argument patterns, such as Toulmin's argument pattern (1958), then they can clearly recognize and distinguish the elements and the structure of an argument, estimate whether the claim can be supported by data, whether it warrants backing and a qualifier or whether a rebuttal is appropriate.

Some researchers have established a few frameworks to evaluate students' competence in writing scientific argumentation. For example, Lee *et al.* (2014) described different levels of students' competence in written scientific argumentation on the basis of Toulmin's argument pattern (1958). Another example of an evaluating framework is Ryu and Sandoval's framework (2012). The rubric they used was adapted from earlier efforts to assess the epistemic criteria (Sandoval and Millwood, 2005). They evaluated the students' task of argument construction from four dimensions, namely, casual structure, causal coherence, citing evidence and explicit justification.

Other evaluating frameworks considered some new theoretical foundations except structure. Kelly and Takao's work (2002) considered the epistemic status of students' claims in their writing and sorted these according to the model presented by Latour (1987). This form of analysis allowed for the consideration of claims at multiple levels of theoretical generality and matched well with the categorical description of the transactional use of language in informative writing. In Sandoval's work (2003), he attended to the conceptual and epistemic aspects of students' scientific explanations. In 2005, Sandoval and Millwood (2005) analyzed the rhetorical reference of students' written expressions. The analysis of rhetorical reference therefore aimed at how students refer to those inscriptions in their explanations. A number of recent research projects examined the impact of argumentation on conceptual understanding in science, some frameworks concentrated on students' conceptual understanding in written arguments, such as those presented by Aydeniz *et al.* (2013), Venville and Dawson (2010).

Aim of this study

In China there has been a strong tradition in science education of acquiring scientific knowledge rather than carrying out scientific practice, thus, no research has yet been undertaken to evaluate the competence of Chinese students' in scientific argumentation, and science teachers in China seldom guide students participating in scientific argumentation explicitly. However, understanding the average performance level of Chinese students in scientific argumentation tasks would be of great significance in Chinese science education research and practice, it would be helpful for Chinese teachers to understand what the Chinese students' current levels of scientific argumentation are, and to design curriculum materials and teaching strategies based on these levels. This paper presents a study which aimed at evaluating 578 (304 males and 274 females) Chinese students' competence of participating in five scientific argumentation tasks with a chemistry background.

From the review of the frameworks for evaluating students' competence of written scientific argumentation, in this study of evaluation, at least four dimensions needed to be considered, they were structure components, logic for justification, content quality and language. The dimension of structure components and logic for justification were constructed on the basis of informal logic, which reflected the logical context of scientific argumentation. The dimension of content quality embodied the conceptual understanding of scientific argumentation. The rhetorical features of scientific argumentation could be diagnosed under the dimension of language.

Based on the item response theory, the Rasch model was applied to produce one scale between the items and the subjects, to produce data consistent with the Rasch model, and to validate the measurement. The scores from the Rasch analysis were used to reveal the results of the study.

Methods

In this study, both qualitative and quantitative methods were used. Firstly, some written scientific argumentation tasks for evaluation were constructed, and the participants in this study completed these tasks. For the data collection, firstly, the qualitative method was used to code the students' answers, and based on the framework for tool design (Table 3), the content criteria and performance criteria for quantitative data acquirement were designed. When analyzing the quantitative data, a qualitative method was also used to explain the results.

Evaluation of tasks

The tasks for written scientific argumentation were designed for a chemistry background. Firstly, five topics, namely chemical reactions, elements and compounds, the structure of substances, chemical experiments and the issue of social science in chemistry, were chosen as the topics for the five tasks.

Table 1 Basic information and key evaluation points for the five tasks

Name	Topic	Characteristic	Key evaluation points
Where is the fog?	Elements and compounds	There are three different claims and a description of experimental facts in the task.	1. Select one claim; 2. State the evidence supporting the selected claim; 3. State the warrants connecting evidence and claim correctly; 4. Construct rebuttals.
Is there any chemical reactions?	Chemical reaction	The question is a yes/no question and the data is presented in data table.	1. Construct one claim; 2. State the evidence supporting the selected claim; 3. State the warrants connecting evidence and claim correctly.
How to choose desiccant?	Chemical experiment	There are three different claims and a data table presenting data in the task.	1. Select one claim; 2. State the evidence supporting the selected claim; 3. State the warrants connecting evidence and claim correctly; 4. Construct rebuttals.
How many types of crystal water exist?	The structure of substance	The question is an open question and the data is presented in a data table.	1. Construct one claim; 2. State the evidence supporting the selected claim; 3. State the warrants connecting evidence and claim correctly.
Can we increase the tax of gasoline?	Social science issue (SSI) in chemistry	The question is a yes/no question and the data of history and experiment is presented in graph.	1. Construct one claim; 2. State the evidence supporting the selected claim; 3. State the warrants connecting evidence and claim correctly.

All of the tasks were presented as tests. The contents or backgrounds of these five tests were derived from tests from a Chinese college entrance examination, PISA, or some other researchers' evaluation instruments, for example, those presented in McNeill's paper (2009). According to the threshold model of content knowledge transfer (Sadler and Fowler, 2006), which proposed two knowledge thresholds around which the quality of argumentation could reasonably be expected to increase, the knowledge acquired by the students was taken into consideration when the topics were chosen, otherwise they would not have been able to finish the written scientific argumentation tasks successfully as they would not have had the required level of understanding of the topics under discussion. In this study, our aim was to investigate the competence of students from different grades in written scientific argumentation, therefore, the tasks set were based on chemistry knowledge which should have been familiar to the youngest students.

For each task, the background was first presented. This was followed by a core question, the students were instructed to construct a claim to solve this question, or choose one claim from the possible claims designed by the researchers. Tests for which there were different possible claims, were suitable for evaluating students' competence in refuting. The students were then presented with some useful data or experimental facts to help them construct evidence. These data could be presented in the form of words, or as a table or graph. All the tasks are presented in Appendix 1.

After the tasks were designed, suggestions were obtained from five experts in science education (include two professors majoring in science education in one normal university in Wuhan and three professional science teachers in high schools in Wuhan) for revisions relating to the accuracy of the content of each task. When the revisions had been completed, the final version of the evaluation task was completely constructed. The basic information and the key evaluation points of these tasks are shown in Table 1.

Participants

In China, the middle school includes grade 7, grade 8 and grade 9, high school includes grade 10, grade 11 and grade 12. But in the Chinese curriculum system, chemistry courses are taught from grade 9 to grade 12. In this study, our aim was to find out what Chinese students' competence is generally like, and to diagnose the differences in the students' competence between the different genders, grades (especially grade 9 to 12) and school levels. The answers to these questions are required for further research as Chinese chemistry teachers seldom teach students how to present scientific argumentation. We hoped to find out what the performance of Chinese students' participating in scientific argumentation would be in the context of chemistry teaching in China which is concentrated too much on knowledge learning.

However, as mentioned before, when students participate in scientific argumentation, they must be equipped with enough chemistry knowledge (Sadler and Fowler, 2006). Thus, when the participants were chosen to represent each grade level, the students were selected at the end of each grade rather than at any other time period. So, the best time for evaluation was the beginning of the new academic year (in China, a whole academic year is from September to the next August). The evaluation tasks were given to participants on the 9th Sept., 2014. At this time, students in China enter a new grade, but the students have not yet been taught any new chemistry topics in the new grade, so these students were chosen to be representatives of the former completed grade.

Although we could easily choose new students from grade 10, grade 11 and grade 12 to represent the completed grade levels of 9, 10 and 11, respectively, it was not easy to choose students to represent grade 12 as the students had already left school and started college. To solve this problem, some freshmen from university were selected to represent students who had just finished grade 12. There were some reasons for this. Firstly, before 9th Sept., 2014, they had been taught all of the Chinese chemistry courses taught in the four years of miiddle school and high school, but had not had any formal tuition in chemistry at college because they were receiving military training in the period of our study (the time period prior to the military training was the summer vacation following completion of grade 12), so their level of chemistry knowledge was consistent with the previously mentioned requirements of this study, namely they had completed grade 12.

In order to make sure that the students chosen were representative of the average level of Chinese students in each grade, the sampling procedure was also taken into account. The high school students (grade 10 to grade 12) participating in the study were from Wuhan, Chengdu, Guangzhou and Rizhao. The four cities in which the high schools are located are in different geographical regions of China, and were randomly chosen. When considering the statistical analysis of the results, it was helpful to eliminate the differences between the different areas. Next, we chose students from five high schools to participate. In this study, these five schools were classified into three groups, they were A-level schools (high schools in Guangzhou and Rizhao), B-level schools (two high schools in Wuhan) and C-level schools (high school in Chengdu). In China, high schools are always divided, by the government and by society, into the key school of the province, the key school of the city and the general school. In China, the standards of classification are the achievement of the students, such as the percentage entering university, the number of awards they have received from the government or the society, the social assessment of society, and so on. This information can be easily found from the website of each high school, and every high school states which category they belong to clearly in their website. For example, from the website of the high school in Guangzhou which was selected for this study, it can be seen that this school was honored as the key school of Guangdong province in the 1970s.

The freshmen were from a normal university located in Wuhan. This normal university is a key comprehensive university directly under the administration of the Chinese Ministry of Education. The freshmen who participated in our study all majored in chemistry, but they had not been taught learned any college chemistry courses in the study period as we mentioned before. The chemistry major of this university enrolls new students from almost all of the provinces in China every year, and from different school levels in each province. In order to keep correspondence with the high school students, which may represent the average level of this grade, we randomly selected participants from all freshmen in the college of chemistry, so that the samples could also eliminate the differences between different areas in China. Then, these freshmen were classified into three categories based on the level of their high schools (level A to C). Two methods were used for classification, firstly, before testing, every freshman had to write the name of his(her) high school and state which category it belonged to. Secondly, the researcher carried out an internet search of the name of each high school and checked the corresponding websites to make sure that the classification was correct.

The numbers of students who participated in the first and second pilot tests were 120 and 140. The number of students who participated in the formal test was 578 (304 males and 274 females). Basic data on the students participating in the formal test are presented in Table 2.

All the tests were held as an additional task beside the formal school curriculum in order to evaluate the students' competence in written scientific argumentation in the context of chemistry, and every test was implemented during their self-study time per week. Before every test, information about the aims of this study was stated clearly by the researchers, and also printed on the test paper. Furthermore, all the participants were told that this was only a study about the competence of Chinese students, and the information about their name, their age, their school would always remain secret, and the results of the test would not influence their study in the future. So the participants would agree to take part in this test, and they would not be under any pressure.

Tool

The tool for evaluation was constructed based on existing research mentioned in the theoretical framework on evaluation criteria of the competence of written scientific argumentation. As shown in Table 1, in total there were 17 key evaluation points for all of the tasks, so there should also have been 17 evaluation items. For each evaluation item, the level of the students'

Table 2 Basic data on the students participating in the formal test

School level in this study	City	Total number		Grade 9	Grade 10	Grade 11	Grade 12		
							A level	B level	C level
A	Guangzhou	Male	48	18	15	15	—	—	—
		Female	50	16	17	17	—	—	—
		Total	98	34	32	32	—	—	—
	Rizhao	Male	54	13	18	23	—	—	—
		Female	46	19	15	12	—	—	—
		Total	100	32	33	35	—	—	—
B	Wuhan	Male	61	20	21	20	—	—	—
		Female	34	12	9	13	—	—	—
		Total	95	32	30	33	—	—	—
	Wuhan	Male	60	18	21	21	—	—	—
		Female	39	15	13	11	—	—	—
		Total	99	33	34	32	—	—	—
C	Chengdu	Male	51	15	15	21	—	—	—
		Female	41	15	13	13	—	—	—
		Total	92	30	28	34	—	—	—
—	Wuhan (College students)	Male	30	—	—	—	7	8	15
		Female	64	—	—	—	16	18	30
		Total	94	—	—	—	23	26	45
Total			578	161	157	166	94		

Table 3 The framework for the tool design

Structure components	Level 1	Level 2	Level 3	Level 4
Claim	No claim	The claim is scientific	—	—
Evidence	No evidence	The evidence is scientific	The evidence are scientific and sufficient	Using detailed, precise and unambiguous language based on level 3
Warrant	No warrant	The warrant is scientific	The warrants are scientific and sufficient	Using detailed, precise and unambiguous language based on level 3
Rebuttal	No rebuttal	Refuting one claim effectively	Refuting all claims effectively	Using detailed, precise and unambiguous language based on level 3

performance also needed to be determined. The first level we assessed concerned the structure components of an argument, in other words whether the student could put forward the corresponding structure component for each evaluation item, such as claim, evidence, warrant and rebuttal. The second level concerned the content quality, it was necessary for the student not only to put forward the corresponding structure component of an argument, but also to use a scientific structure component. The third level was about the logic of justification, for example, the evidence and warrant should be sufficient, even though they were all scientific. The highest level was about language, the student had to use detailed, precise and unambiguous language in expressing an argument.

These were general criteria for evaluation, but there also existed some particular cases. When the evaluation item was about constructing or selecting one claim, only structure components and content quality had been considered because there was no problem with the logic of justification, and the expression of the claim was simple enough, such as "Alice!", "The chemical reaction happened!" (because of these theoretical considerations, levels 3 and 4 of the claim were not required). For evaluating the competence of refuting (tasks 1 and 3), as both claims needed to be refuted, it was assumed that level 2 was about refuting one claim effectively and level 3 was about refuting all of the claims effectively. A rebuttal was deemed to be scientific when a student could refute a claim effectively, and the evidence for refuting a claim was sufficient and the warrants were reasonable. The highest level of rebuttal was also about the language use.

The framework for the tool design is presented in Table 3.

When designing the concrete tool, both content criteria and performance criteria were included in the whole criteria. The content criteria included a description of each evaluation item and level, and the performance criteria gave the specific description which students could show in written scientific argumentation tasks. When coding the answers of students in the first and second pilot tests, the content criteria and the performance criteria of each evaluation item were constructed, in particular, some examples of how student responses belonged to different levels were illustrated. In Table 4 are presented the content criteria and performance criteria for evaluating the quality of evidence in task 1 (evaluation item: 02T1E). The first two letters represent the number of evaluation items, the third and forth letters represent the task number which the specific evaluation item belongs to, the last letter represents the structure components of an argument which are related to the specific evaluation item. C represents Claim, E representsd Evidence, W represents Warrant, R represents Rebuttal.

There is a question that the content criteria and performance criteria mentioned in Table 4 were not completely in accordance with the framework for the tool design, such as the sufficiency of evidence was not concerned. This is reasonable because the framework only reflects the general evaluation criteria, in the design of the concrete criteria, how the framework is applied depends on the task and the coding results of the two pilot tests. In task 1, no other evidence could be used to support Grace's claim, and no students put forward other evidence to support this claim, therefore, insufficient evidence should have been put forward to support this claim.

The Rasch model was used to modify and improve the tool based on the data from the first and second pilot tests. Winsteps Version 3.72.0 was applied to calculate the ability estimate for each student and the item difficulty estimate for each item.

Table 5 shows the estimations of student competence and item difficulty, and some indicators calculated from the data in the two pilot tests and in the formal test. It is obvious that some indicators were always in accordance with the Rasch model (MNSQ $\approx$ 1.00 and ZSTD $\approx$ 0.00, see Bond and Fox, 2007). The person separation increased from the first pilot test to the formal test as well as the reliability, which means that the discrimination and the reliability of the tool are more suitable for this study.

Bubble charts (Fig. 1) were used to represent the fitness to the Rasch model of each evaluation item. If a bubble was

Table 4 The content criteria and performance criteria for evaluating the quality of evidence in task 1 (evaluation item: 02T1E)

Content criteria	Performance criteria
Level 1 (0 point): no evidence, or evidence which could not support Grace's claim.	No statement about evidence, or some statements like "The volatility of NH_3 is stronger than HCl".
Level 2 (1 point): put forward evidence which could support Grace's claim.	Based on the data in task, showed the evidence like "In the same time, the displacement of NH_3 is longer than HCl, so the rate of NH_3 is also stronger than HCl".
Level 3 (2 points): put forward evidence which could support Grace's claim and the language of evidence description was detailed, precise and unambiguous.	The language of evidence description was detailed, precise and unambiguous, for example, there were expressions like "in the same time", "the white ring is near the cotton wool soaked with concentrated hydrochloric acid" in the argument.

Table 5 The estimations and some indicators in three tests

Test		Estimation	Error	Infit		Outfit		Separation	Reliability
				MNSQ	ZSTD	MNSQ	ZSTD		
First pilot test	Student competence	−1.24	0.44	1.01	0.0	0.97	0.0	1.28	0.62
	Item difficulty	0.00	0.18	1.00	0.1	0.97	−0.1	4.67	0.96
Second pilot test	Student competence	−1.03	0.58	1.01	0.0	0.97	0.0	1.54	0.70
	Item difficulty	0.00	0.20	1.00	0.1	0.97	0.0	7.54	0.98
Formal test	Student competence	−0.94	0.57	1.01	0.0	0.96	0.0	1.67	0.74
	Item difficulty	0.00	0.10	1.00	0.0	0.96	−0.1	15.94	1.00

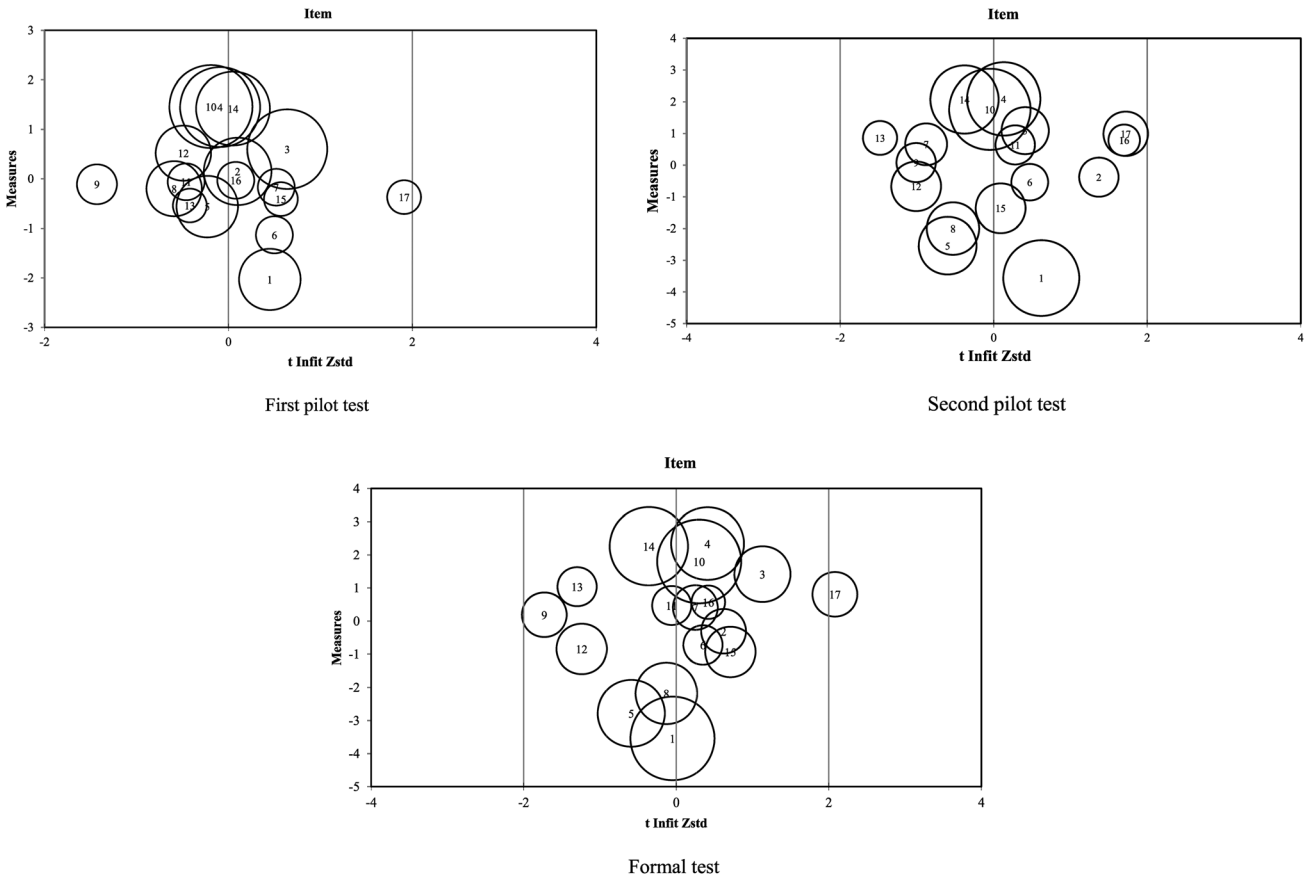


Fig. 1 Bubble charts for the three tests.

located at [−2, 2] of the *X* axis, the results obtained for the evaluation item were deemed to be suitable (according to Bond and Fox, 2007, the number in the center of the bubble indicates the number of the evaluation item). From Fig. 1, it is obvious that almost all of the bubbles were located at [−2, 2], so almost all of the evaluation items were fit according to the Rasch model, except for 17T5W. As this measurement was not a high risk test, according to Huang (2012), the criteria could be accepted.

The dimensionality maps (Fig. 2) for the three tests are also shown. If the letter was located at [−0.4, 0.4] of the *Y* axis, or the letter was near to this interval, the results for the evaluation item were deemed to satisfy the assumption of unidimensionality, or became more unidimensional (according to Bond and Fox, 2007, the letter indicates the number of evaluation items, the corresponding relation can be found in Winsteps Version 3.72.0 according to the report on data analysis). All three dimensionality maps indicated that the modified and improved tool had been suitably adapted for the research because more letters were located in a suitable interval.

The formal criteria are presented in Appendix 2. The reason why the tool had become more suitable for this study was because some of the criteria used for evaluation had been modified. For example, for the items used to evaluate the students' competence in rebuttal, level 2 and level 3 were all defined as refuting (a) claim(s) effectively. When this item was modifed, the results showed that few students could achieve the highest level, for which they needed to use detailed, precise and unambiguous language to express their rebuttal, so level 4 of these items needed to be modified, or needed to be incorporated into other levels. But there was still a problem relating to the evaluation of

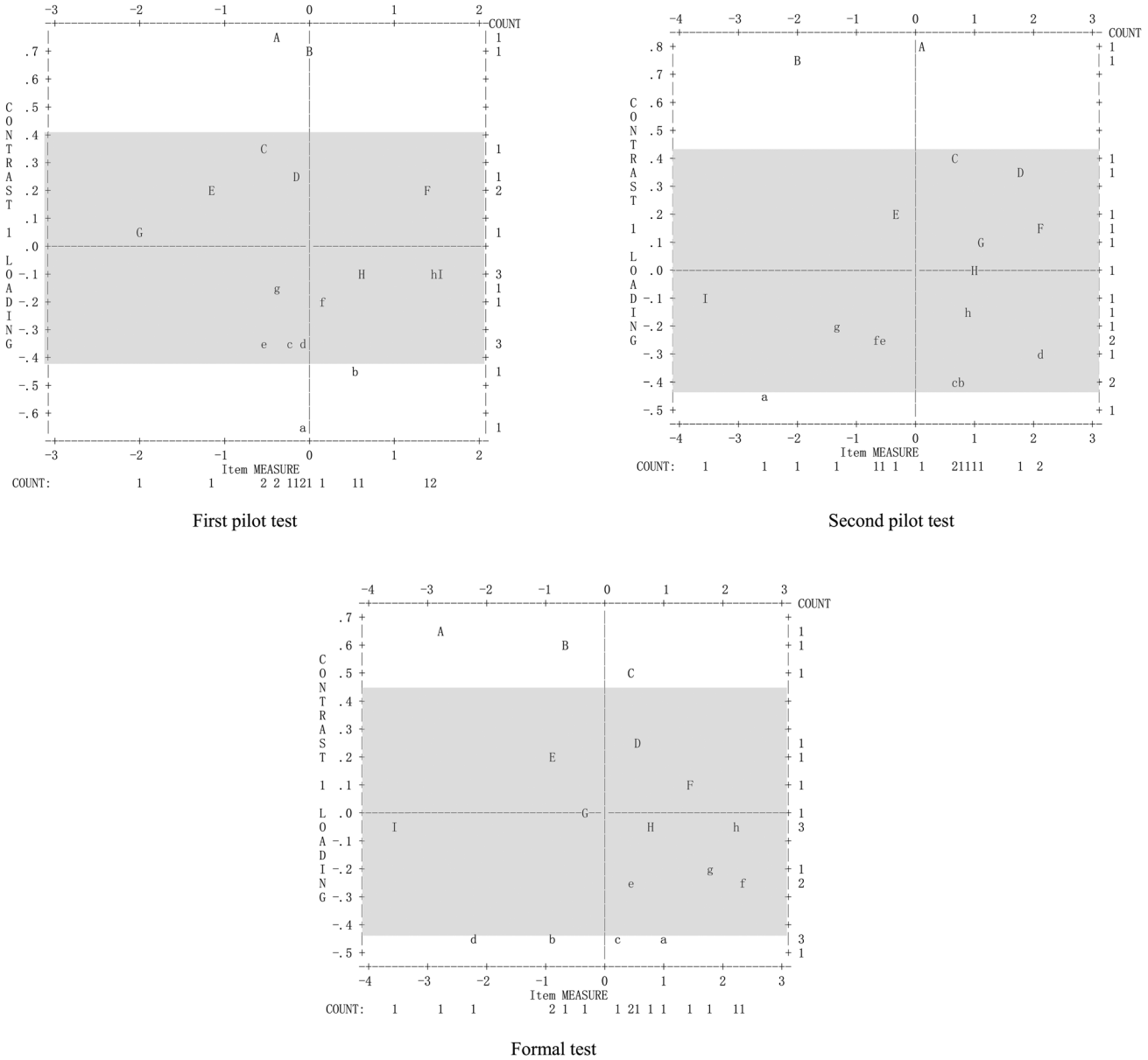


Fig. 2 Dimensionality maps for the three tests.

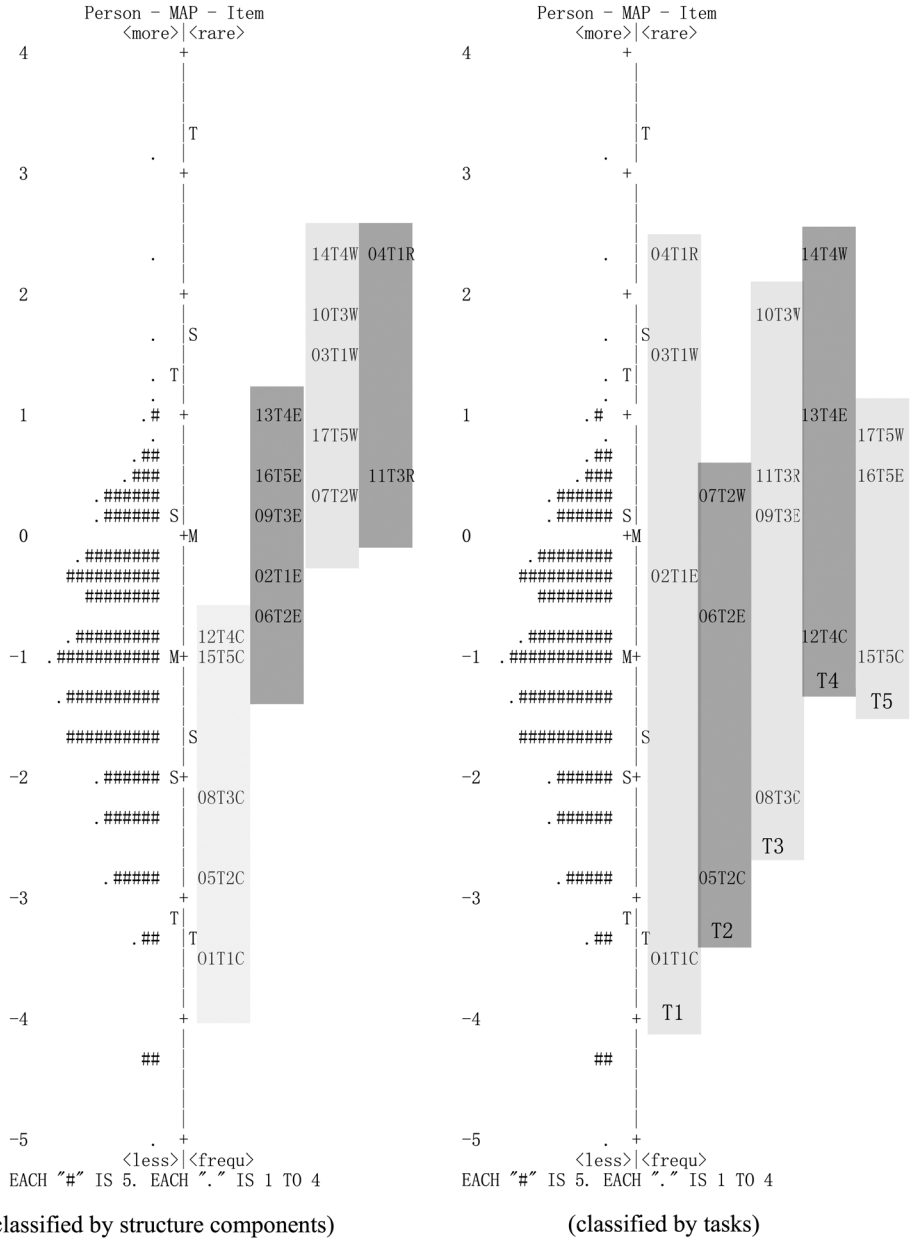
the students' competence in the use of language? The word "effectively" in the definition of levels 2 and 3 could express the meaning that the statements of rebuttal should not only be scientific and logical, but also had to be detailed, precise and unambiguous. So before the second pilot test, the content criteria and the performance criteria of item 4 and item 11 were both revised.

Results

General descriptions

As is shown in Table 5, in the formal test, the estimation of student competence using the Rasch model (−0.94) was lower than that of item difficulty (0.00). This indicated that Chinese students' competence in written scientific argumentation was generally weak.

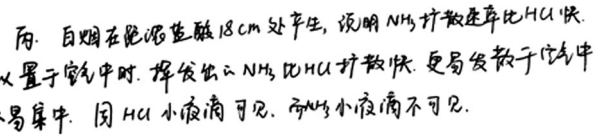
Fig. 3 shows the Wright map of the formal test and is based on the Rasch model. It is easy to understand that students could put forward claims more easily than evidence. On the other hand, stating warrants and rebuttals for an argument seemed more difficult. The task itself was a key factor which influenced the students' competence in written scientific argumentation. Specifically, the difficulties of task 2, task 3 and task 4 increased progressively, and the corresponding difficulties of making claims, providing evidence and warrants for these tasks also increased. Although for task 1, the difficulty of putting forward claims, providing evidence and warrants followed the same trends as for tasks 2, 3 and 4, but the gaps between them were much bigger. For task 5, as it was a task about written scientific argumentation in SSI, there were some differences in the results obtained compared to those obtained for the other tasks.



(classified by structure components) (classified by tasks)

Fig. 3 Wright map for the formal test.

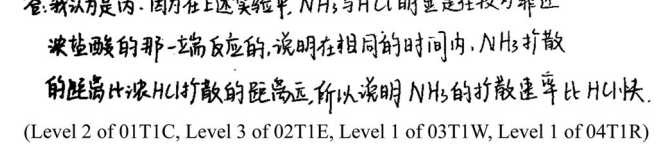
More specifically, based on students' answers, for task 1, most students could put forward a correct claim, but as the experiment only showed the diffusion distances of NH_3 and HCl, and their claims were about the diffusion rates of NH_3 and HCl, they could not state the evidence completely, such as "in the same time".



(Level 2 of 01T1C, Level 2 of 02T1E, Level 2 of 03T1W, Level 1 of 04T1R)

(Translation: Grace: A white ring formed 18 cm from the cotton wool which was soaked with concentrated hydrochloric acid, so the diffusion rate of NH_3 is faster than that of HCl. When in air, NH_3 diffuses faster, and could not be concentrated, so the small drop of NH_3 cannot be seen.)

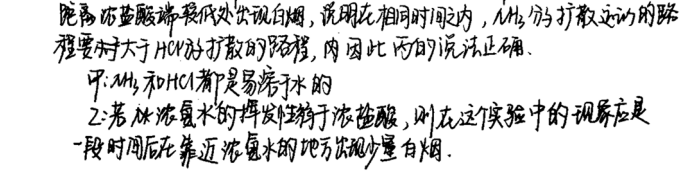
What's worse, fewer students showed the warrants which may connect the claim and evidence, such as "NH_3 diffuses rapidly around, so $NH_3{\cdot}H_2O$ could not be concentrated, so the fog is not obvious." This problem might have arisen because the students only considered Grace's claim and the experimental facts although relevant, were ignored so that they could answer the core questions in task 1.



(Level 2 of 01T1C, Level 3 of 02T1E, Level 1 of 03T1W, Level 1 of 04T1R)

(Translation: I think Grace is correct, because in this experiment, the reaction between NH_3 and HCl began near HCl, which means that at the same time, NH_3 diffuses for a longer period of time, so the diffusion rate of NH3 is stronger than that of HCl.)

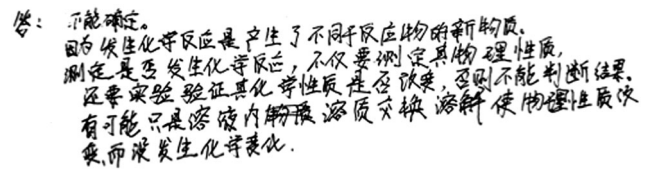
In task 1, if the students wanted to support their claim more effectively, they refuted Alice's and Peter's claims. Few students could put forward rebuttals, as they needed to use their own knowledge, such as the solubility of NH_3 and HCl. Even for rare exceptions, the students could not refute them correctly. For example, one student refuted Peter's claim based on the experimental facts of task 1 because he confused the two concepts of volatility and diffusion.



(Level 2 of 01T1C, Level 3 of 02T1E, Level 1 of 03T1W, Level 1 of 04T1R)

(Translation: In the long glass tube, both pieces of cotton wool were simultaneously inserted at one end of the glass tube, and a white ring was found near the HCl after several minutes. This means that at the same time, the diffusion path of NH_3 is longer than that of HCl, so Grace is correct! By the way, according to Alice, NH_3 and HCl are both soluble substances, and if Peter is right, the white ring would be found near the cotton wool soaked with concentrated ammonia solution.)

Take task 2 as another example. Most students thought that a chemical reaction had occurred, but, a few students misunderstood the nature of the chemical reaction. For example, someone concluded that the characteristics of a chemical reaction were the changes of the chemical properties of substances.

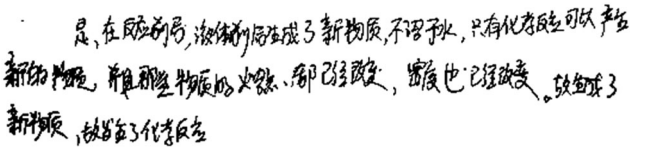


(Level 1 of 05T2C, Level 1 of 06T2E, Level 1 of 07T2W)

(Translation: I am not sure. A chemical reaction will happen when new substances are produced. We should not only measure the physical properties of new substances but also the chemical properties, or we cannot judge whether a chemical reaction has happened. In this table, maybe the solutes changed, so the physical properties have also changed.)

Some students only presented a conclusion about the evidence which could support their correct claim without any data, such as ''new substances were produced''. Others showed some irrelevant data, such as the volume of the reactants and the products. But most of the students, could not put forward sufficient evidence from the data table, such as the changes to the melting points, the boiling points, and the densities of every reactant and product.

For warrants, some students did not show any warrants about whether the chemical reaction happened, the others' warrants were not complete (a complete warrant should include: ① if the melting points, the boiling points and the densities of two substances are different, they are different substances; ② a chemical reaction could have happened if a new substance has been produced.).

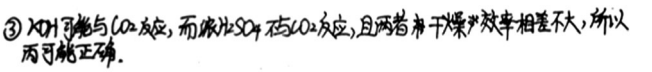


(Level 2 of 05T2C, Level 3 of 06T2E, Level 2 of 07T2W)

(Translation: Yes! Before and after the reaction, new substances were produced because layer C cannot dissolve in water. Only a chemical reaction can produce new substances. And also, the melting points and the densities are all changed, so a chemical reaction has happened.)

The students found task 3 to be more difficult than task 2 in stating claim because in this task they needed not only the information in the task, but also had to rely on some of their own knowledge and understanding (such as what is the positive ion and the negative ion, what kind of substances contain ions, what kind of substances can react with CO_2, *etc.*) If they forgot this basic knowledge of chemistry, they could not provide a correct claim.

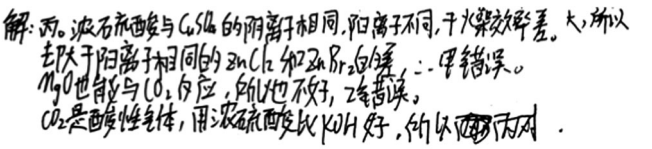
Students who agreed with Grace's claim, provided the correct evidence, but it was not sufficient, for example, the efficiency of drying of H_2SO_4 and KOH are almost same, it would not influence the capability of these two desiccants. If the students showed this evidence, the argumentation might be more reasonable, such as:



(Level 2 of 08T3C, Level 3 of 09T3E, Level 1 of 10T3W, Level 1 of 11T3R)

(Translation: The third one. KOH could react with CO_2 but H_2SO_4 does not, and their efficiency of drying are similar, so Grace may be correct.)

In task 3, as the aim of this task was producing CO_2 in a lab, and the claim was about selecting the desiccant, the data was about the drying efficiency of the desiccant, so a warrant was needed, such as ''a desiccant which can consume CO_2 is not good enough''. But few students were able to mention this. For a rebuttal, the competence of refuting was higher than that for task 1, maybe the students were able to use the information presented in the data table rather than having to rely on their own knowledge and understanding. But still some students could not refute correctly due to some misconceptions.

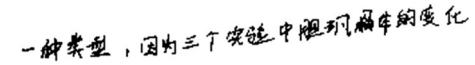


(Level 2 of 08T3C, Level 2 of 09T3E, Level 1 of 10T3W, Level 2 of 11T3R)

(Translation: H_2SO_4 and $CuSO_4$ both consist of SO_4^{2-}, and $ZnCl_2$ and $ZnBr_2$ both consist of Zn^{2+}, but the difference

between H_2SO_4 and $CuSO_4$ is greater than that for the latter two, so the first claim was wrong. MgO can react with CO_2, so the second claim was wrong too. CO_2 is an acidic gas, so H_2SO_4 is more suitable than KOH, so the third one was right.)

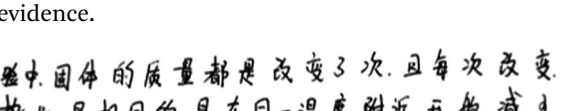
As task 4 was about the structure of a substance, students could use abstract thinking to analyze the macro-phenomena and micro-structure, so it was more difficult. For example, some students could not put forward a correct claim at first.



(Level 1 of 12T4C, Level 1 of 13T4E, Level 1 of 14T4W)

(Translation: One type, because the changing trend of data in these three experiments are the same.)

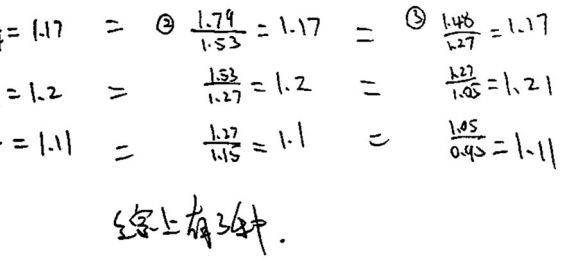
Three types of problems arose when students attempted to provide evidence for a correctly stated claim. The first one was that they could not carry out calculations properly using the data presented in the data table, and only used words to provide evidence.



(Level 2 of 12T4C, Level 2 of 13T4E, Level 1 of 14T4W)

(Translation: In these three experiments,† the mass are all change three times with the same the changing trend, and at the same temperature, so there are 3 types of force.)

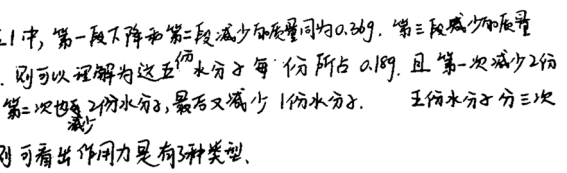
The second problem was the opposite: only data and formulae were presented and no words were used to provide evidence for the claim.



(Level 2 of 12T4C, Level 2 of 13T4E, Level 1 of 14T4W)

(Translation: 3 types)

The third problem was that some students did not provide sufficient evidence. They only showed one group of data for one experiment and ignored the remaining data.

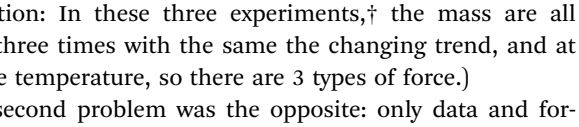


(Level 2 of 12T4C, Level 2 of 13T4E, Level 1 of 14T4W)

(Translation: In the first experiment,‡ the mass reduces 0.36 g at the first and second time, and 0.18 g at the third time. It can be calculated that the mass of one H_2O in 2.50 g $CuSO_4$ is 0.18 g. So in 2 H_2O lose at the first and second time, and 1 H_2O at the last time. That means there are three types of force between five H_2O and Cu^{2+} in $CuSO_4{\cdot}5H_2O$.)

A great number of students found the warrant item for task 4 difficult because they needed to understand the nature of the forces between particles, and the relationship of forces, energy, and temperature. So few students could achieve the highest level of 14T4W.

Task 5 was used to evaluate the students' competence in written scientific argumentation in the context of a social science issue. The students had to put forward a claim about a social science issue, they always came up with a claim from either a social aspect or a scientific aspect and provided evidence for their claim, but they did not include both aspects of the issue. Many students thought that CO_2 was the main reason for the greenhouse effect, but when the students were asked to further consider whether the government should raise the tax of petrol to forbid citizens from driving, they only pointed out that "this method is palliative", "the biggest source of CO_2 is from industry", or "it is inconvenient to people", which only related to social reasons. Some of the students only showed some active specific actionable recommendations about how to slow down the carbon emission without answering the question about whether the government should raise the tax of petrol.



(Level 1 of 15T5C, Level 1 of 16T5E, Level 1 of 17T5W)

(Translation: We should develop new technologies and new methods to reduce the emission of CO_2, not raise the tax of petrol!)

From the results above, it was further demonstrated that the task itself was a key factor which influenced the students' competence in written scientific argumentation, this might be due to differences in the background knowledge required for each task, the thinking styles required to solve the problem, and the information provided for each task or the students own knowledge which could be used for argumentation.

Possible influence factors

The first factor we studied, which might have influenced the students' competence in written scientific argumentation, was gender.

Table 6 shows the results of *t*-tests for the mean score in the Rasch model for the variables of gender. From the results presented it is indicated that there were no significant differences between the males and females in terms of the total number of points for the students' competence in written scientific argumentation. For two specific items (05T2C, 15T5C), there were significant differences, but no regular findings about which gender might perform better could be obtained, this

† All experiments mentioned in task 4, see Appendix 1.

‡ The first experiment mentioned in task 4, see Appendix 1.

Table 6 The results of t-tests for the variables of gender

	Gender	Number	Mean	SD	t
Total points	Male	304	−0.935	1.242	−0.061
	Female	274	−0.929	1.057	
01T1C	Male	304	−2.650	0.647	0.971
	Female	274	−2.705	0.717	
02T1E	Male	304	−0.892	1.642	−1.674
	Female	274	−0.661	1.665	
03T1W	Male	304	0.096	0.888	1.354
	Female	274	0.002	0.779	
04T1R	Male	304	0.577	0.725	0.251
	Female	274	0.562	0.734	
05T2C	Male	304	−2.251	0.689	1.982*
	Female	274	−2.373	0.791	
06T2E	Male	304	−0.793	1.665	1.229
	Female	274	−0.962	1.647	
07T2W	Male	304	−0.478	1.431	1.923
	Female	274	−0.705	1.406	
08T3C	Male	304	−1.652	0.966	0.277
	Female	274	−1.674	0.979	
09T3E	Male	304	−0.613	1.600	0.764
	Female	274	−0.710	1.438	
10T3W	Male	304	0.924	0.667	0.901
	Female	274	0.877	0.599	
11T3R	Male	304	−0.483	1.275	0.362
	Female	274	−0.520	1.173	
12T4C	Male	304	−0.855	1.102	0.192
	Female	274	−0.872	1.102	
13T4E	Male	304	−0.584	1.268	0.052
	Female	274	−0.589	1.209	
14T4W	Male	304	0.701	0.571	−0.128
	Female	274	0.707	0.565	
15T5C	Male	304	−1.024	1.098	−2.526*
	Female	274	−0.794	1.094	
16T5E	Male	304	−0.629	1.091	−1.364
	Female	274	−0.502	1.138	
17T5W	Male	304	−0.440	1.548	0.633
	Female	274	−0.519	1.458	

*$p < 0.05$.

might be attributed to an error in sampling. To sum up, the factor of gender was not found to influence the students' competence in written scientific argumentation.

Next, the factor of grade was also studied. There were 4 groups of grades in the participants of this study, so one-way ANOVA had to be used to check the differences between the grades. The results are presented in Tables 7–9. Table 7 shows the results for a homogeneity test of variances, Table 8 shows the descriptive statistics for the total points and each item for the different ages. Table 9 is a summary of the one-way ANOVA. To compare the differences between the four grades, the Scheffe method was used to check the items that satisfied the homogeneity of variance assumptions. For other items, the Tamhane's T2 method was used. In these 3 tables, all the statistics used were based on the Rasch analysis.

From the results presented in Tables 7–9, it can be seen that the students' competence in written scientific argumentation for these 4 grades was significantly different. It can be seen that the total number of points for students in grade 9 were the lowest, grade 10, grade 12 and grade 11 followed in turn. There were significant differences between the total number points for students in grade 9 and grade 10, also for students in grade 10 and grade 11. Although the total number points for students in grade 12 were lower than those in grade 11, there were no significant differences. This indicated that the students' competence in written scientific argumentation increased from grade from 9 to 11, and decreases a little in grade 12, but the decrease was not obvious.

For a more detailed analysis, firstly, for the different tasks, the students' competence in putting forward claims were significantly different. For tasks 2, 3 and 4, the students' competence in putting forward claims increased from grade 9 to grade 11, but were constant from grades 11 to grade 12 except for task 2 (the latter was significantly weaker than the former). For task 1, there were no significant differences between each of the grades, but for task 5, the competence of grade 9 students' in putting forward claims was significantly stronger than the next 3 grades.

Secondly, the students' competence in putting forward evidence showed almost the same characteristics for each claim. Some differences could be seen for task 2, in which the competence of students in grade 12 was still weaker than that of grade 11 students, but the difference was not significant. For task 5, the competence of grade 9 students' for putting forward evidence was stronger than that of students in grade 10, and the competence of grade 11 students' was stronger than that of students in grade 12. The difference between students in grade 10 and grade 11 was significant, the latter was stronger.

Thirdly, for putting forward warrants, there was very little difference between the four grades for tasks 1 and 3. For task 2, the competence of grade 9 students was significantly weaker than that of students in grade 11, weaker than that of students in grade 10 and grade 12, and the competence of grade 12 students was also weaker than that of students in grade 10 and grade 11. For task 4, the competence of grade 11 students was the strongest, and was significant. For task 5, the competence of grade 9 students was weaker than that of students in grade 10 and grade 12, significantly weaker than that of students in grade 11. Grade 12 students' competence was also weaker than that of students in grades 10 and 11. Generally speaking, the competence of putting forward warrants did not increase for students in higher grades.

Table 7 The results for a homogeneity test of variances (differences between grades)

Item	Levene statistic	Significance	Item	Levene statistic	Significance	Item	Levene statistic	Significance
Total points	0.779	0.506	06T2E	2.370	0.070	12T4C	19.861	0.000
01T1C	1.759	0.154	07T2W	0.614	0.606	13T4E	23.484	0.000
02T1E	1.824	0.142	08T3C	25.482	0.000	14T4W	18.111	0.000
03T1W	1.121	0.340	09T3E	1.293	0.276	15T5C	4.874	0.002
04T1R	26.573	0.000	10T3W	5.840	0.001	16T5E	6.534	0.000
05T2C	11.584	0.000	11T3R	14.443	0.000	17T5W	2.857	0.036

Table 8 Descriptive statistics for the total points and each item for different grades

Item	Grade	Mean	SD	Item	Grade	Mean	SD	Item	Grade	Mean	SD
Total points	9	−1.327	1.175	06T2E	9	−1.026	1.690	12T4C	9	−1.352	0.976
	10	−1.063	1.204		10	−0.861	1.654		10	−0.819	1.103
	11	−0.583	1.034		11	−0.616	1.564		11	−0.562	1.067
	12	−0.654	1.002		12	−1.086	1.729		12	−0.629	1.085
01T1C	9	−2.686	0.695	07T2W	9	−0.878	1.380	13T4E	9	−1.126	0.852
	10	−2.720	0.736		10	−0.471	1.445		10	−0.598	1.203
	11	−2.639	0.633		11	−0.410	1.417		11	−0.218	1.376
	12	−2.651	0.651		12	−0.586	1.411		12	−0.292	1.288
02T1E	9	−0.875	1.584	08T3C	9	−1.886	1.063	14T4W	9	0.638	0.461
	10	−0.857	1.742		10	−1.711	0.998		10	0.651	0.484
	11	−0.733	1.677		11	−1.544	0.900		11	0.843	0.745
	12	−0.587	1.594		12	−1.408	0.787		12	0.659	0.457
03T1W	9	0.083	0.842	09T3E	9	−1.091	1.377	15T5C	9	−0.787	1.094
	10	−0.007	0.837		10	−0.746	1.536		10	−1.203	1.069
	11	0.050	0.842		11	−0.453	1.508		11	−0.917	1.103
	12	0.098	0.839		12	−0.138	1.577		12	−0.649	1.069
04T1R	9	0.423	0.519	10T3W	9	0.823	0.507	16T5E	9	−0.668	1.024
	10	0.504	0.641		10	0.924	0.668		10	−0.785	1.001
	11	0.762	0.931		11	0.912	0.651		11	−0.346	1.251
	12	0.591	0.705		12	0.981	0.738		12	−0.430	1.113
05T2C	9	−2.330	0.759	11T3R	9	−0.801	0.986	17T5W	9	−0.769	1.383
	10	−2.290	0.727		10	−0.655	1.170		10	−0.342	1.493
	11	−2.203	0.640		11	−0.246	1.276		11	−0.326	1.564
	12	−2.491	0.863		12	−0.181	1.439		12	−0.472	1.572

In terms of providing rebuttals, grade 11 students were much better at refuting than students in grade 9 and grade 10. There was no difference in the results obtained for students in grade 11 and grade 12.

The third factor which was studied was that of the school level. There were 3 school levels in the participants of this study, so one-way ANOVA analysis had to be used to check the differences between each school level. The results are presented in Tables 10–12. Table 10 shows the results for the homogeneity test of variances, Table 11 shows the descriptive statistics for the total number of points and each item in different school levels. Table 12 is a summary of the one-way ANOVA. To compare the differences between the three schools, the Scheffe method was used to check the items so that the homogeneity of the variance assumptions was satisfied. For other items, the Tamhane's T2 method was used. In these 3 tables, all the statistics were based on the Rasch analysis.

It was obvious that the students' competence in written scientific argumentation for these 3 kinds of schools were significantly different. After comparison, it was obvious that the total number of points for the students from the A level schools were significantly higher than those from the B and C level schools. Although the total number of points for the students from the B level schools were lower than for those from the C level schools, the difference was not significant. This indicated that, in the A level schools, the students' competence in written scientific argumentation was much stronger, but in B and C level schools, there were no significant differences between the students' competence.

More detailed results from the perspective of each structure component would also be analyzed. In terms of putting forward claims, students from the A level schools were able to raise much more perfect claims for all of the tasks set, than were the other students. But for students from B and C level schools, no significant differences were found between all of the tasks except for task 1. These results were similar to those obtained for stating evidence.

In terms of stating warrants, it was obvious that there were no significant differences between the students from all three level schools, except for task 2, for which the A level school students' competence was significantly stronger than that of the students from the C level schools. These results indicated that when giving warrants, students always had the same difficulties irrespective of the kind of school they attended, or they had all ignored to state the relationships between claims and evidence in their scientific argumentation.

Finally, when refuting in task 3, there were significant differences between the students from levels A and B, and from levels A and C. But this results was not be obtained for task 1.

Conclusions and discussions

This study focused on evaluating Chinese students' competence in written scientific argumentation. On the basis of the results, it might be generally concluded that Chinese students' competence in written scientific argumentation is weak, and is influenced by dozens of factors.

Firstly, students could put forward claims and evidence more easily than warrants and rebuttals. These conclusions were also drawn in previously published research, such as that of Erduran *et al.* (2004), Kelly and Takao (2002), Felton and Kuhn (2001), Sandoval (2003). When the students solved the problem that was presented in each task, firstly, they aimed at giving the answers to the question which was included in each task based on their own knowledge. Of course, as for all of the

Table 9 Summary of one-way ANOVA (differences between the grades)

		Sum of squares	df	Mean square	*F*	*Post hoc* tests	
Total points	Between Groups	55.288	3	18.429	14.757**	9 < 10*	10 < 11**
	Within Groups	716.868	574	1.249		9 < 11**	10 < 12**
	Total	772.156	577			9 < 12**	
01T1C	Between Groups	0.614	3	0.205	0.439	n.s.	
	Within Groups	267.278	574	0.466			
	Total	267.891	577				
02T1E	Between Groups	6.260	3	2.087	0.760	n.s.	
	Within Groups	1575.077	574	2.744			
	Total	1581.337	577				
03T1W	Between Groups	0.899	3	0.300	0.425	n.s.	
	Within Groups	404.983	574	0.706			
	Total	405.882	577				
04T1R	Between Groups	10.324	3	3.441	6.664**	9 < 11**	10 < 11*
	Within Groups	296.411	574	0.516			
	Total	306.734	577				
05T2C	Between Groups	5.109	3	1.703	3.137*	12 < 11*	
	Within Groups	311.647	574	0.543			
	Total	316.756	577				
06T2E	Between Groups	18.993	3	6.331	2.322	n.s.	
	Within Groups	1565.061	574	2.727			
	Total	1584.053	577				
07T2W	Between Groups	21.003	3	7.001	3.504*	9 < 11*	
	Within Groups	1146.819	574	1.998			
	Total	1167.823	577				
08T3C	Between Groups	16.862	3	5.621	6.114**	9 < 11*	10 < 12*
	Within Groups	527.638	574	0.919		9 < 12**	
	Total	544.500	577				
09T3E	Between Groups	63.808	3	21.269	9.554**	9 < 11**	10 < 12*
	Within Groups	1277.883	574	2.226		9 < 12**	
	Total	1341.691	577				
10T3W	Between Groups	1.684	3	0.561	1.393	n.s.	
	Within Groups	231.314	574	0.403			
	Total	232.998	577				
11T3R	Between Groups	38.564	3	12.855	8.889**	9 < 11**	10 < 11*
	Within Groups	830.061	574	1.446		9 < 12**	10 < 12*
	Total	868.625	577				
12T4C	Between Groups	59.071	3	19.690	17.660**	9 < 10**	
	Within Groups	640.007	574	1.115		9 < 11**	
	Total	699.079	577			9 < 12**	
13T4E	Between Groups	77.559	3	25.853	18.352**	9 < 10**	
	Within Groups	808.594	574	1.409		9 < 11**	
	Total	886.152	577			9 < 12**	
14T4W	Between Groups	4.533	3	1.511	4.777**	9 < 11*	10 < 11*
	Within Groups	181.565	574	0.316			
	Total	186.098	577				
15T5C	Between Groups	22.346	3	7.449	6.316**	10 < 9**	
	Within Groups	676.900	574	1.179		10 < 12**	
	Total	699.246	577				
16T5E	Between Groups	18.934	3	6.311	5.195**	10 < 11**	
	Within Groups	697.417	574	1.215			
	Total	716.351	577				
17T5W	Between Groups	20.333	3	6.778	3.023*	9 < 11*	
	Within Groups	1287.038	574	2.242			
	Total	1307.371	577				

n.s.; $p > 0.05$; *$p < 0.05$; **$p < 0.01$.

tasks data which may be useful was provided, the students naturally presented the relative data or evidence to support their claims. So students would state the claims and evidence easily as long as they could solve these questions successfully. But when providing evidence, the students sometimes did not realize the relevance and sufficiency of the evidence, which sometimes lead to the phenomenon that students provided evidence that was a bit too complex for their claims. These conclusions were also drawn by McNeill and Krajcik (2007). But, why were the students not able to show the warrants and rebuttals successfully? As Kelly and Takao (2002) mentioned, for some students, they could not state the relationship between a claim and evidence clearly, and they always tended to strengthen their own claims, without taking into consideration the claims of other people, particularly some counter-claims, or counter-argumentations (Pontecorvo and Girardet, 1993). This might be due to a lack of basic epistemology of scientific practices, especially scientific argumentation (Sandoval and Millwood, 2007),

Table 10 Results for the homogeneity test of variances (differences between the school levels)

Item	Levene statistic	Significance	Item	Levene statistic	Significance	Item	Levene statistic	Significance
Total points	3.274	0.039	06T2E	1.549	0.213	12T4C	2.028	0.132
01T1C	25.000	0.000	07T2W	2.599	0.075	13T4E	0.115	0.891
02T1E	20.144	0.000	08T3C	13.569	0.000	14T4W	5.220	0.006
03T1W	6.977	0.001	09T3E	1.136	0.322	15T5C	0.951	0.387
04T1R	3.312	0.037	10T3W	0.587	0.556	16T5E	12.862	0.000
05T2C	17.154	0.000	11T3R	1.737	0.177	17T5W	1.898	0.151

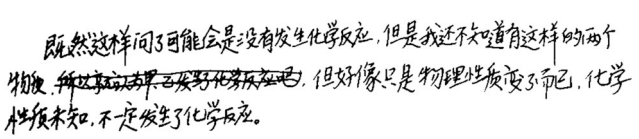
Table 11 Descriptive statistics for the total number of points and each item in the different school levels

Item	School	Mean	SD	Item	School	Mean	SD	Item	School	Mean	SD
Total points	A	−0.608	1.046	06T2E	A	−0.676	1.674	12T4C	A	−0.656	1.087
	B	−1.145	1.206		B	−1.011	1.568		B	−1.050	1.082
	C	−1.113	1.136		C	−0.971	1.747		C	−0.896	1.103
01T1C	A	−2.599	0.571	07T2W	A	−0.377	1.507	13T4E	A	−0.440	1.221
	B	−2.640	0.634		B	−0.644	1.371		B	−0.724	1.273
	C	−2.858	0.866		C	−0.827	1.323		C	−0.600	1.195
02T1E	A	−0.451	1.481	08T3C	A	−1.548	0.902	14T4W	A	0.741	0.636
	B	−0.960	1.715		B	−1.790	1.031		B	0.703	0.570
	C	−1.031	1.748		C	−1.642	0.963		C	0.644	0.430
03T1W	A	−0.037	0.772	09T3E	A	−0.218	1.561	15T5C	A	−0.816	1.097
	B	0.116	0.883		B	−1.049	1.386		B	−1.000	1.100
	C	0.092	0.863		C	−0.744	1.503		C	−0.938	1.104
04T1R	A	0.613	0.821	10T3W	A	0.889	0.618	16T5E	A	−0.302	1.251
	B	0.559	0.695		B	0.920	0.662		B	−0.741	0.956
	C	0.517	0.617		C	0.893	0.624		C	−0.723	1.039
05T2C	A	−2.221	0.659	11T3R	A	−0.287	1.239	17T5W	A	−0.459	1.592
	B	−2.300	0.735		B	-0.645	1.186		B	−0.512	1.449
	C	−2.465	0.848		C	−0.615	1.230		C	−0.452	1.460

for example, they only of thought evidence that could a support claim, and did not treat the claim and the evidence as two independent facts, they did not understand how to use a warrant to connect the evidence to a claim. When arguing, they never considered a situation in which the claim was untenable, they seldom took rhetoric into consideration, they did not recognize persuasion, and only stressed their own claims.

Secondly, it was obvious that the task themselves influences the performance of students in scientific argumentation, and some demographic variables might have also been factors which influenced the students' competence in scientific argumentation. In this study, we studied three demographic variables, they were gender, grade level and school level. From the results, gender was not found to be a factor which could influence this competence, but the grade level and school level were more important. For the variable of grade level, we found that students' competence in written scientific argumentation in different grades were significantly different, this could be explained by using the threshold model of content knowledge transfer which was constructed by Sadler and Fowler (2006), but there might be little difference. The results verified the conclusion clearly that when students' knowledge increases (as students' move to more senior grades), the argumentation quality could reasonably be expected to increase, but knowledge was only one factor which might lead to the thresholds. Some other factors like cognitive ability, thinking style might also affect argumentation. So it was preferred to say it was a threshold model of scientific cognition transfer, rather than treating it as a threshold model of content knowledge transfer.

But what was the explanation for the competence of grade 12 students' in written scientific argumentation being weaker than that of students in grade 11, but not significant? And for 05T2C, the former students being significantly weaker than the later in putting forward a claim? This might be an interesting question but we did not have sufficient data to solve it. Maybe, some grade 12 students did not provide a scientific claim for task 2 because they subconsciously thought that the knowledge needed for solving the problem was too simple because they had already learned about the concept of ''chemical reaction'' in middle school, and they thought that the task designers would investigate some particular knowledge which they had forgotten, for example, one grade 12 student reamrked:



(Translation: There was no chemical reactions since you asked me this question, but I do not know whether these two substances exist, ~~so may be there was a chemical reaction,~~ but it seems only the physical properties changed. I do not know about the chemical properties, so I am not sure whether there were any chemical reactions.)

Table 12 Summary of the one-way ANOVA (differences between the school levels)

		Sum of squares	df	Mean square	*F*	*Post hoc* tests
Total points	Between groups	37.719	2	18.859	14.765**	A > B**
	Within groups	734.438	575	1.277		A > C**
	Total	772.156	577			
01T1C	Between groups	6.1	2	3.05	6.699**	A > C**
	Within groups	261.791	575	0.455		B > C*
	Total	267.891	577			
02T1E	Between groups	39.663	2	19.832	7.397**	A > B**
	Within groups	1541.674	575	2.681		A > C**
	Total	1581.337	577			
03T1W	Between groups	2.851	2	1.425	2.034	n.s.
	Within groups	403.031	575	0.701		
	Total	405.882	577			
04T1R	Between groups	0.812	2	0.406	0.763	n.s.
	Within groups	305.922	575	0.532		
	Total	306.734	577			
05T2C	Between groups	5.056	2	2.528	4.663**	A > C*
	Within groups	311.7	575	0.542		
	Total	316.756	577			
06T2E	Between groups	14.122	2	7.061	2.586	n.s.
	Within groups	1569.932	575	2.73		
	Total	1584.053	577			
07T2W	Between groups	18.355	2	9.177	4.591**	A > C*
	Within groups	1149.468	575	1.999		
	Total	1167.823	577			
08T3C	Between groups	6.537	2	3.269	3.494*	A > C*
	Within groups	537.963	575	0.936		
	Total	544.5	577			
09T3E	Between groups	77.508	2	38.754	17.627**	A > B**
	Within groups	1264.183	575	2.199		A > C**
	Total	1341.691	577			
10T3W	Between groups	0.12	2	0.06	0.148	n.s.
	Within groups	232.879	575	0.405		
	Total	232.998	577			
11T3R	Between groups	16.531	2	8.265	5.578**	A > B**
	Within groups	852.094	575	1.482		A > C*
	Total	868.625	577			
12T4C	Between groups	17.329	2	8.664	7.308**	A > B**
	Within groups	681.75	575	1.186		
	Total	699.079	577			
13T4E	Between groups	8.937	2	4.468	2.929	n.s.
	Within groups	877.215	575	1.526		
	Total	886.152	577			
14T4W	Between groups	0.808	2	0.404	1.253	n.s.
	Within groups	185.291	575	0.322		
	Total	186.098	577			
15T5C	Between groups	3.849	2	1.925	1.591	n.s.
	Within groups	695.397	575	1.209		
	Total	699.246	577			
16T5E	Between groups	25.527	2	12.764	10.624**	A > B **
	Within groups	690.824	575	1.201		A > C**
	Total	716.351	577			
17T5W	Between groups	0.428	2	0.214	0.094	n.s.
	Within groups	1306.943	575	2.273		
	Total	1307.371	577			

n.s.; $p > 0.05$; *$p < 0.05$; **$p < 0.01$.

When stating warrants and rebuttals, there were no significant differences between the different grades, this might due to the fact that the difficulties for reasoning and refuting were greater. Of course, some students could not distinguish between evidence and claims, and could not reason logically using evidence and claims (Kuhn, 1989), so they simply thought that only the answer was enough.

The significant differences in the competence in written scientific argumentation for students' from A level schools and those from B or C schools were also obvious. It could be easily concluded that students from much higher level schools could put forward more scientific claims and evidence for all of the tasks, as they had greater knowledge and understanding, and a stronger ability for problem solving. But when reasoning and refuting, the situations were different. We might conclude that students who had a higher level of understanding of knowledge, skills and abilities could do well in putting forward claims and evidence, but when reasoning and refuting, there

might be no differences between other kinds of students. This conclusion should remind us that when teaching students to construct a relationship between claims and evidence, they must pay close attention to others and this is most important in teaching argumentation.

This study has provided some insight into the Chinese chemistry curriculum. Firstly, chemistry teachers should pay more attention to the values of scientific argumentation, and help students to participate in different kinds of scientific argumentation activities. It is essential for teachers to design some instructional activities (such as debate competition) to help students starting controversies based on students' own ideas about chemical phenomena. Maybe some ideas have come about as a result of their misconceptions, but this is also helpful, and arguing can lead to conceptual change. For instance, when students argue about the structure of NaCl, some students may hold the opinion that one ionic bond belongs to one Na^+ and one Cl^- without any regard for the crystalline structure of NaCl. This may be one of the main misconceptions of Chinese students, especially in grade 11. If teachers were to organize students with proper claims and students with misconceptions into one group to model what the real structure of NaCl is based on the data of its melting point, density and so on, it would be helpful for the development of the students' competence in scientific argumentation as well as for achieving conceptual change. It is crucial for Chinese science teachers to present their students with examples of what a good argument is (Hogan and Maglienti, 2001), and also to instruct them in some skills for persuading so that they could improve their skills in effective scientific argumentation. Also, when designing the tasks for scientific argumentation, the difficulties, the categories and the structures of different tasks should be considered, especially when the background is a social science issue. Chemistry teachers should help students to analyze the questions from a scientific aspect and from a social aspect, logically and reasonably. In order to reach this aim, Chinese teachers should try to integrate chemistry knowledge with some other subjects in chemistry teaching. The most striking examples of this in the Chinese chemistry curriculum are some topics related to the chemical industries such as the ammonia industry, the chlorine alkali industry and the sulfuric acid industry which can be taught well with comprehensive backgrounds including a of knowledge of geography, economics, demography, *etc.*

In addition to the strategies that teachers should focus on for the teaching of skills for scientific argumentation, from the results presented above it can also be concluded that when teaching students in different grade levels, the chemistry teachers should teach them how to participate in scientific argumentation differently. Higher level students (such as students in higher grades or in A/B level schools), are equipped with a lot of knowledge to solve chemistry problems, but sometimes they may take unnecessary pains to study an insignificant problem, or are interrupted by some misconceptions when they are thinking. For such students, chemistry practice is more important than the knowledge itself. The teachers should help students to realize what the aim is of the problem solving task, what methods should be applied, and how the students should cooperate with each other effectively. Sometimes teachers should offer a timely reminder, or point out their mistakes. For students in the lower levels, such as students in lower grades or in B/C level schools, the teachers should take into account the level of the students' knowledge and understandings, maybe the teachers can firstly help them to look back on what they have learned before, then ask them to participate in scientific argumentation. A good example of this can be the ethyl acetate synthesis. If the argumentation task is about how to synthesize ethyl acetate more effectively (faster with a higher conversion rate), for higher level students, the teachers could ask them to think about what kinds of phenomena can represent the aim of ''effective'', and how to achieve this aim by applying different chemical methods. However, for the lower level students, reviewing the knowledge of the basic principle of synthesis, especially the Le Chatelier's principle at first may be much more valuable.

The third suggestion is that chemistry teachers should help students to put forward warrants to link claims and evidences in scientific argumentation, and ask them to refute others or to include some counter-argumentations. As we know, when taking part in scientific argumentation, everyone should reason logically from evidence to claims, and consider whether the evidence can support the claims. But some students cannot always express this process, sometimes they even ignore this process, and regard evidence and claims as equal (Kuhn, 1989). To realize the different functions of evidence and warrants is helpful, Chinese students should pay more attention to the structure of scientific argumentation. Supposing arguing about the comparison of the acidity of H_2SO_4 and H_3PO_4. Some students may state that ''the number of non-hydroxy-O in a molecule of H_2SO_4 is higher, so it is a stronger acid''. In this argument, only the claim (it is a stronger acid) and evidence (the number of non-hydroxy-O in a molecule of H_2SO_4 is higher) have been stated. If the teachers point out that adding a warrant like ''based on Pauling's empirical rules that the more non-hydroxy-O the oxoacid has, the stronger the acidity would be'', the claim would be more confirmed. Besides, when arguing, Chinese students may prefer to listen to the argumentation of others, but not refute or reconsider the argumentation of others, especially the counter-claim or counter-argument. Some researchers have also pointed out that students seldom refute the claims of others, or challenge others directly which is the main problem in their scientific argumentation (Felton and Kuhn, 2001). Therefore, how to teach students to become critics is a world-wide science education issue. In the case above, teachers can ask to compare the acidity of H_3PO_4, H_3PO_3 and H_3PO_2, and find out that the fact (data of pK_a, *etc.*) may be different from the result based on Pauling's empirical rules. This is a good exception and useful for criticizing, and also, it helps students to understand the rules much better than counting the number of non-hydroxy-O which focuses on the structure of the molecule.

Appendix 1: the evaluation tasks

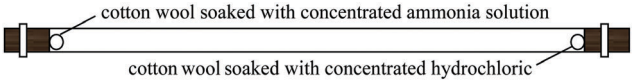
Task 1: After opening the reagent bottle of concentrated ammonia solution and concentrated hydrochloric acid, it is easy to notice the phenomenon that the fog is appeared above the reagent bottle of concentrated hydrochloric acid rather than concentrated ammonia solution. Three students put forward their ideas that may explain this phenomenon.

Alice: NH_3 could not easily combine with water but HCl can do so.

Peter: The volatility of concentrated ammonia solution is weaker than concentrated hydrochloric acid.

Grace: The diffusion rate of NH_3 is stronger than HCl.

In order to know whose explanation may be correct, the teacher showed them an experiment. Firstly, he soaked one cotton wool in the concentrated ammonia solution and a next piece in the concentrated hydrochloric acid. Next, he inserted both cotton wools simultaneously at one end of the glass tube (l = 50 cm, $\varnothing$ = 2 cm) and the other end of it respectively and then quickly inserted a rubber bungs at both ends of the tube as shown below. When the two gases reacted, there was a white ring formed 18 cm from the cotton wool soaked with concentrated hydrochloric acid and 32 cm from the other one.



Now, please state whose idea you will support and why.

Task 2: Carlos wants to know whether two liquids will react with each other. He uses an eye-dropper to get a sample from the two liquids A and B. He takes some measurements of each of the two samples. Then he stirs the two liquids together and heats them. After stirring and heating the liquids, they form two separate layers: layers C and D. Carlos uses an eye-dropper to get a sample from each layer. He takes some measurements of each sample. Here are his results:

		Measurements			
Data		Melting point (°C)	Volume (cm^3)	Solubility in water	Density ($g\ cm^{-3}$)
Before stirring and heating	Liquid A	−7.9	2.00	Yes	0.96
	Liquid B	−89.5	2.00	Yes	0.81
After stirring and heating	Layer C	−91.5	2.00	No	0.87
	Layer D	0.01	2.00	Yes	0.99

Based on these data, please help Carlos to judge whether a chemical reaction occurred when Carlos stirred and heated A and B and why.

Task 3: The capability of a desiccant can be measured by the efficiency of drying (the mass of water vapor left after drying in 1 m^3). The table below shows the efficiency of drying of different desiccants.

Desiccant	Efficiency of drying	Desiccant	Efficiency of drying
MgO	0.008	H_2SO_4	0.003
CaO	0.200	$CuSO_4$	1.400
$ZnCl_2$	0.800	KOH	0.002
$ZnBr_2$	1.100	NaOH	0.160

When producing CO_2 in lab, three students put forward their ideas about how to dry CO_2.

Alice: The efficiency of drying may relate to the property of ion, and the negative ion has more effects to the efficiency of drying than the positive ion.

Peter: MgO is more suitable than CaO to dry CO_2.

Grace: H_2SO_4 is more suitable than KOH to dry CO_2.

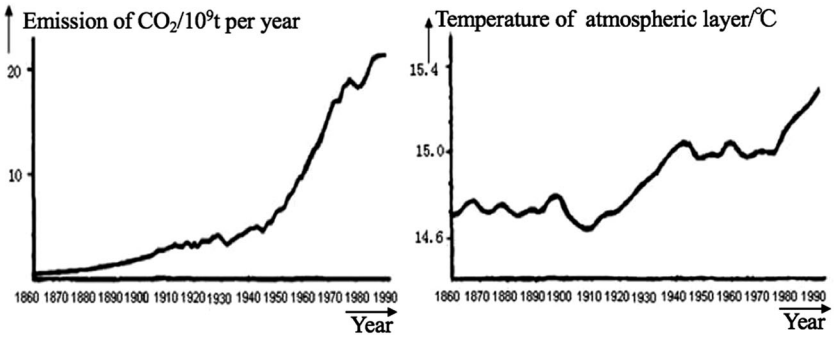
Now, please state whose idea you will support and why.

Task 4: Someone think that in $CuSO_4{\cdot}5H_2O$, the types of force between five H_2O and Cu^{2+} are the same, but others think there are different types of force. In order to test, Lucy and her classmates heated 2.50 g $CuSO_4{\cdot}5H_2O$, 1.79 g $CuSO_4{\cdot}5H_2O$, and 1.48 g $CuSO_4{\cdot}5H_2O$ separately, and weighed the mass left of each sample after dehydration. The three groups of experiment data is shown in the following table.

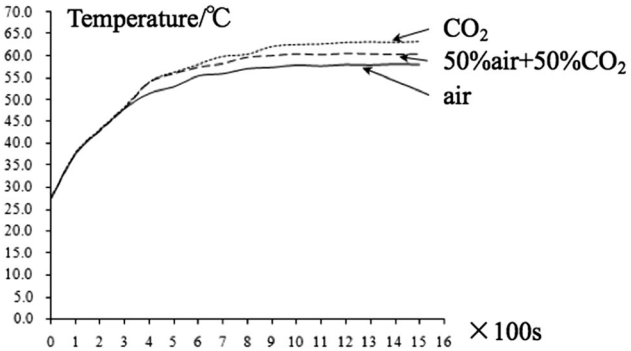
Temperature (°C)		25	25–104	106–114	116–258	260–280
Experiment 1	m_1/g	2.50	2.50 ± 0.01	2.14 ± 0.01	1.78 ± 0.01	1.60 ± 0.01
Experiment 2	m_2/g	1.79	1.79 ± 0.01	1.53 ± 0.01	1.27 ± 0.01	1.15 ± 0.01
Experiment 3	m_3/g	1.48	1.48 ± 0.01	1.27 ± 0.01	1.05 ± 0.01	0.95 ± 0.01

Based on these data, how many types of force between five H_2O and Cu^{2+} in $CuSO_4{\cdot}5H_2O$? why?

Task 5: The greenhouse effect become more and more serious in recent years. One day, Lily read an article in newspaper who claimed that one of the main reasons for the temperature increasing is the increased emission of CO_2. In order to understand this issue in detail, she found two graphs in the library.



She also used three infra-red lamps to shine on a bottle of air, a bottle of CO_2 and a bottle of mixed air which contain 50% CO_2 and 50% air separately (the light intensities and the distances between each infra-red lamps and bottle are all the same). She found that the temperature of these three bottles change differently and she used a graph to show this differences.



Based on this data, Lily thought the government should raise the tax of petrol to forbid people driving, so that the emission of CO_2 would be decrease. Do you agree with her? Why?

Appendix 2: the evaluation criteria

01T1C

Content criteria	Performance criteria
Level 1 (0 point): no claim, or not select Grace's claim.	No statement about claim, or select Alice's or Peter's claim.
Level 2 (1 point): select Grace's claim.	Select Grace's claim explicitly.

02T1E

Content criteria	Performance criteria
Level 1 (0 point): no evidence, or evidence which could not support Grace's claim.	No statement about evidence, or some statements could not support Grace's claim like ''the volatility of NH_3 is stronger than HCl''.
Level 2 (1 point): put forward scientific evidence which could support Grace's claim.	Based on the data in task, show the scientific evidence like ''in the same time, the displacement of NH_3 is longer than HCl, so the rate of NH_3 is also stronger than HCl''.
Level 3 (2 points): put forward evidence which could support Grace's claim and the language of evidence description was detailed, precise and unambiguous.	The language of evidence description was detailed, precise and unambiguous, for example, there were expressions like ''in the same time'', ''the white ring is near the cotton wool soaked with concentrated hydrochloric acid'' in the argument.

03T1W

Content criteria	Performance criteria
Level 1 (0 point): no warrant, or warrants which could not connect scientific claim and evidence.	No statement about warrant, or some statements of warrant are not suitable, like ''the drop of $NH_3{\cdot}H_2O$ is such small to be seen''.
Level 2 (1 point): put forward suitable warrant which could connect scientific claim and evidence.	Put forward suitable warrant such as ''NH_3 diffuses faster, could not be concentrated, so the small drop of NH_3 cannot be seen''.
Level 3 (2 points): put forward suitable warrant which could connect scientific claim and evidence and the language of warrant description was detailed, precise and unambiguous.	The language of suitable warrant description was detailed, precise and unambiguous, for example, there are expressions like ''small drop of $NH_3{\cdot}H_2O$'', not ''small drop of NH_3''.

04T1R

Content criteria	Performance criteria
Level 1 (0 point): no rebuttal, or refute Grace's claim, or the rebuttals of Alice's or Peter's claim were not correct.	No statement about rebuttal, or state some evidence and warrants which did not support Grace's claim, or the rebuttals of Alice's or Peter's claim were not correct, such as ''For Peter, if he were right, the white ring would be found near the cotton soaked with concentrated ammonia solution.''
Level 2 (1 point): based on evidence and warrants, refute only one claim of Alice's or Peter's correct.	Based on evidence and warrants, refute only one claim of Alice's or Peter's correct. To refute Alice's claim, the statement might be as ''the solubility of NH_3 is bigger than HCl'' (evidence). To refute Peter's claim, the statement might be that ''the volatility is influenced by temperature'' (warrant).
Level 3 (2 point): based on evidence and warrants, refute both Alice's and Peter's claim correct.	State both two rebuttals in Level 2.

05T2C

Content criteria	Performance criteria
Level 1 (0 point): no claim, or state wrong claims.	No statement about claim, or state wrong claims such as ''there were no chemical reactions''.
Level 2 (1 point): state the claim like ''the chemical reaction happened''.	State the claim like ''the chemical reaction happened'' explicitly.

06T2E

Content criteria	Performance criteria
Level 1 (0 point): no evidence, or evidence which could not support the claim ''the chemical reaction happened''.	No statement about evidence, or some statements could not support scientific claim like ''the volumes of substances were not change''.
Level 2 (1 point): put forward scientific evidence which could support the claim ''the chemical reaction happened'', but not sufficient.	Based on the data in task, put forward only one scientific evidence of ''the melting points changed'', ''the densities changed'', and ''the solubility changed''.
Level 3 (2 points): put forward sufficient evidence which could support the claim ''the chemical reaction happened'' and the language of evidence description was detailed, precise and unambiguous.	Put forward all scientific evidence mentioned in level 2 and the language of evidence description was detailed, precise and unambiguous, for example, there were expressions like ''the melting points, the densities changed and the solubility are all changed'' in the argument.

07T2W

Content criteria	Performance criteria
Level 1 (0 point): no warrant, or warrants which could not connect scientific claim and evidence.	No statement about warrant, or some statements of warrant are not suitable, like ''layer C and layer D are not chemical substances''.
Level 2 (1 point): put forward suitable warrant which could connect scientific claim and evidence, but not sufficient.	Put forward only one suitable warrant of ''a chemical reaction happened when new substances had been produced'', and ''different substances have different properties''.
Level 3 (2 points): put forward suitable warrant which could connect scientific claim and evidence and the language of warrant description was detailed, precise and unambiguous.	Put forward all suitable warrant mentioned in level 2 and the language of warrant description was detailed, precise and unambiguous.

08T3C

Content criteria	Performance criteria
Level 1 (0 point): no claim, or not select Grace's claim.	No statement about claim, or select Alice's or Peter's claim.
Level 2 (1 point): select Grace's claim.	Select Grace's claim explicitly.

09T3E

Content criteria	Performance criteria
Level 1 (0 point): no evidence, or evidence which could not support Grace's claim.	No statement about evidence, or some statements could not support Grace's claim such as "the efficiency of drying of H_2SO_4 is bigger than KOH."
Level 2 (1 point): put forward scientific evidence which could support Grace's claim, but not sufficient.	Based on the data in task, put forward only one scientific evidence of "the efficiency of drying of H_2SO_4 and KOH are almost the same", and "KOH could react with CO_2".
Level 3 (2 points): put forward sufficient evidence which could support Grace's claim and the language of evidence description was detailed, precise and unambiguous.	Put forward all scientific evidence mentioned in level 2 and the language of evidence description was detailed, precise and unambiguous.

10T3W

Content criteria	Performance criteria
Level 1 (0 point): no warrant, or warrants which could not connect scientific claim and evidence.	No statement about warrant, or some statements of warrant are not suitable, like "H_2SO_4 is in liquid".
Level 2 (1 point): put forward suitable warrant which could connect scientific claim and evidence and the language of warrant description was detailed, precise and unambiguous.	Put forward suitable warrant of "the desiccant which can consume CO_2 is not good enough", and the language of warrant description was detailed, precise and unambiguous.

11T3R

Content criteria	Performance criteria
Level 1 (0 point): no rebuttal, or refute Grace's claim, or the rebuttals of Alice's or Peter's claim were not correct.	No statement about rebuttal, or state some evidence and warrants which did not support Grace's claim, or the rebuttals of Alice's or Peter's claim were not correct, such as "H_2SO_4 and $CuSO_4$ are all consist of SO_4^{2-}".
Level 2 (1 point): based on evidence and warrants, refute only one claim of Alice's or Peter's correct.	Based on evidence and warrants, refute only one claim of Alice's or Peter's correct. To refute Alice's claim, the statement might be as "the efficiency of drying of CaO is 25 times bigger than MgO, but their positive ions are different. The efficiency of drying of $ZnBr_2$ is only 1.375 times bigger than $ZnCl_2$, but their negative ions are different. So negative ion has little influence to the efficiency of drying" (evidence). To refute Peter's claim, the statement might be that "MgO and CaO can react with CO_2, and consume CO_2" (warrant).
Level 3 (2 point): based on evidence and warrants, refute both Alice's and Peter's claim correct.	State both two rebuttals in Level 2.

12T4C

Content criteria	Performance criteria
Level 1 (0 point): no claim, or state wrong claims.	No statement about claim, or state wrong claims.
Level 2 (1 point): state the claim like "there are 3 types of force between five H_2O and Cu^{2+} in $CuSO_4 \cdot 5H_2O$".	State the claim like "there are 3 types of force between five H_2O and Cu^{2+} in $CuSO_4 \cdot 5H_2O$.

13T4E

Content criteria	Performance criteria
Level 1 (0 point): no evidence, or evidence which could not support the claim "there are 3 types of force between five H_2O and Cu^{2+} in $CuSO_4 \cdot 5H_2O$".	No statement about evidence, or some statements could not support scientific claim like "the mass are all change three times with the same the changing trend, and at the same temperature".
Level 2 (1 point): put forward scientific evidence which could support the claim "there are 3 types of force between five H_2O and Cu^{2+} in $CuSO_4 \cdot 5H_2O$", but not sufficient.	Based on the data in task, show the evidence in one or two experiment, such as "in experiment 1, 2.50 g $CuSO_4 \cdot 5H_2O$ loses H_2O in 104 °C, 114 °C and 258 °C, and the mass of losing H_2O is 14.4%, 14.4% and 7.2% of the mass of $CuSO_4 \cdot 5H_2O$. After calculate, when heating, 1 mol $CuSO_4 \cdot 5H_2O$ loses 2 mol, 2 mol and 1 mol H_2O per time".
Level 3 (2 points): put forward sufficient evidence which could support the claim "there are 3 types of force between five H_2O and Cu^{2+} in $CuSO_4 \cdot 5H_2O$".	Based on the data in task, show all evidence in the three evidence to support the claim "there are 3 types of force between five H_2O and Cu^{2+} in $CuSO_4 \cdot 5H_2O$".
Level 4 (3 points): put forward sufficient evidence which could support the claim "there are 3 types of force between five H_2O and Cu^{2+} in $CuSO_4 \cdot 5H_2O$" and the language of evidence description was detailed, precise and unambiguous.	Put forward all scientific evidence mentioned in level 3 and the language of evidence description was detailed, precise and unambiguous (with detail process of calculating).

14T4W

Content criteria	Performance criteria
Level 1 (0 point): no warrant, or warrants which could not connect scientific claim and evidence.	No statement about warrant, or some statements of warrant are not suitable, like ''the mass of $CuSO_4{\cdot}5H_2O$ decreases is due to water vaporization''.
Level 2 (1 point): put forward suitable warrant which could connect scientific claim and evidence.	Put forward suitable warrant of ''the force between five H_2O and Cu^{2+} in $CuSO_4{\cdot}5H_2O$ are different, some force is weaker, so the H_2O which interact with Cu^{2+} by weaker force will lose in lower temperature when heating''.
Level 3 (2 point): put forward suitable warrant which could connect scientific claim and evidence and the language of warrant description was detailed, precise and unambiguous.	Put forward suitable warrant mentioned in level 2 and the language of warrant description was detailed, precise and unambiguous.

15T5C

Content criteria	Performance criteria
Level 1 (0 point): no claim, or state claims which not on the basis of ''CO_2 is the main reason of the green house effect''.	No statement about claim, or not state ''CO_2 is the main reason of the green house effect''.
Level 2 (1 point): based on ''CO_2 is the main reason of the green house effect'', claim whether the government should raise the tax of petrol to forbidden citizens' driving.	Claim whether the government should raise the tax of petrol to forbidden citizens' driving, and state that ''CO_2 is the main reason of the green house effect''.

16T5E

Content criteria	Performance criteria
Level 1 (0 point): no evidence, or evidence which could not support the claim ''CO_2 is the main reason of the green house effect''.	No statement about evidence, or some statements could not support scientific claim like ''CO_2 is the main reason of the green house effect'' like ''sometimes the temperature was fall down''.
Level 2 (1 point): put forward scientific evidence which could support the claim ''CO_2 is the main reason of the green house effect'', but not sufficient.	Based on the data in task, show one evidence like ''when the concentration of CO_2 increases, it can absorb more heat form the sun to keep the earth warm'', or ''the increasing of the concentration of CO_2 is relate to the increasing of temperature of earth''.
Level 3 (2 points): put forward sufficient evidence which could support the claim ''CO_2 is the main reason of the green house effect''.	Based on the data in task, show all evidence mentioned in level 2 to support the claim ''CO_2 is the main reason of the green house effect''.
Level 4 (3 points): put forward sufficient evidence which could support the claim ''CO_2 is the main reason of the green house effect and the language of evidence description was detailed, precise and unambiguous.	Put forward all scientific evidence mentioned in level 3 and the language of evidence description was detailed, precise and unambiguous. (State clearly there is a correlation between to variables.)

17T5W

Content criteria	Performance criteria
Level 1 (0 point): no warrant, or warrants which could not be relate to the claim.	No statement about warrant, or some statements of warrant are not suitable, like ''I am a poor man, I cannot afford the tax of petrol''.
Level 2 (1 point): put forward suitable warrant, but not sufficient.	Put forward one suitable warrant like ''it would have obstructed the development of transportation'', or ''the CO_2 is mainly produced by factories'' and so on to support the claim ''the government should not raise the tax of petrol to forbidden citizens' driving'', or put forward one suitable warrant like ''the burning of petrol would produce more CO_2'' and so on to support the claim ''the government should raise the tax of petrol to forbidden citizens' driving''
Level 3 (2 point): put forward suitable and sufficient warrants and evidence and the language of warrant description was detailed, precise and unambiguous.	Put forward all suitable warrants in one case mentioned in level 2 and the language of warrant description was detailed, precise and unambiguous.

Acknowledgements

We would like to acknowledge the support of the self-determined research funds of Central China Normal University from the colleges' basic research and operation of MOE (CCNU16A05001).

References

Aydeniz *et al.*, (2013), Argumentation and students' conceptual understanding of properties and behaviors of gases, *Int. J. Sci. Math. Educ.*, **10**(6), 1303–1324.

Berland L. K. and Reiser B. J., (2011), Classroom communities' adaptations of the practice of scientific argumentation, *Sci. Educ.*, **95**, 191–216.

Bond T. G. and Fox C. M., (2007), *Applying the Rasch model: fundamental measurement in the human sciences*, 2nd edn, London: Lawrence Erlbaum Associates.

Browne M. N. and Keeley S. M., (1998), *Asking the right questions: a guide to critical thinking*, Prentice Hall.

Cavagnetto A. and Hand B. M., (2012), The importance of embedding argument within science classrooms, in Khine M. S. (ed.), *Perspectives on scientific argumentation: theory, practice and research*, Dordrecht: Springer.

Driver R., Newton P. and Osborne J., (2000), Establishing the norms of scientific argumentation in classrooms, *Sci. Educ.*, **84**, 287–312.

Erduran S., (2007), Methodological foundations in the study of argumentation in science classrooms, in Erduran S. and Jiménez-Aleixandre M. P. (ed.), *Argumentation in science education: perspectives from classroom-based research*, Dordrecht: Springer.

Erduran S., Simon S. and Osborne J., (2004), TAPping into argumentation: developments in the application of Toulmin's argument pattern for studying science discourse, *Sci. Educ.*, **88**, 915–933.

Felton M. and Kuhn D., (2001), The development of argumentative discourse skill, *Discourse Process*, **32**(2&3), 135–153.

Foong C.-C. and Daniel E., (2012), Students' argumentation skills across two socio-scientific issues in a Confucian classroom: is transfer possible? *Int. J. Sci. Educ.*, **34**(1), 1–25.

Garcia-Mila M. and Andersen C., (2007), Cognitive foundations of learning argumentation, in Erduran S. and Jiménez-Aleixandre M. P. (ed.), *Argumentation in science education: perspectives from classroom-based research*, Dordrecht: Springer.

Hogan K. and Maglienti M., (2001), Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions, *J. Res. Sci. Teach.*, **38**, 663–687.

Hong Z. R., Lin H. S., Wang H. H., Chen H. T. and Yang K. K., (2013), Promoting and scaffolding elementary school students' attitudes toward science and argumentation through a science and society intervention, *Int. J. Sci. Educ.*, **35**(10), 1625–1648.

Huang Q., (2012), *Chemistry learning and scientific reasoning: a study on middle school students*, Doctoral dissertation, Beijing Normal University.

Jiménez-Aleixandre M. P. and Erduran S., (2007), Argumentation in science education: an overview, in Erduran S. and Jiménez-Aleixandre M. P. (ed.), *Argumentation in science education: perspectives from classroom-based research*, Dordrecht: Springer.

Kelly G. J. and Takao A., (2002), Epistemic levels in argument: an analysis of university oceanography students' use of evidence in writing, *Sci. Educ.*, **86**, 314–342.

Kuhn D., (1989), Children and adults as intuitive scientists, *Psychol. Rev.*, **96**(4), 674–689.

Latour B., (1987), *Science in action: how to follow scientists and engineers through society*, Cambridge, MA: Harvard University Press.

Lee H. S., Liu O. L., Pallant A., Roohr K. C., Pryputniewicz S. and Buck Z. E., (2014), Assessment of uncertainty-infused scientific argumentation. *J. Res. Sci. Teach.*, **51**(5), 581–605.

Li X., (2004), *Research on scientific rhetoric*, Taiyuan: Shanxi University.

McNeill K. L., (2009), Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena, *Sci. Educ.*, **93**, 233–268.

McNeill K. L., (2011), Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year, *J. Res. Sci. Teach.*, **48**(7), 793–823.

McNeill K. L. and Krajcik J., (2007), Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations, in Lovett M. C. and Shah P. (ed.), *Thinking with data: the Proceedings of the 33rd Carnegie Symposium on Cognition*, Mahwah, NJ: Erlbaum.

Mendonca P. C. and Justi R., (2014), An instrument for analyzing arguments produced in modeling-based chemistry lessons, *J. Res. Sci. Teach.*, **51**(2), 192–218.

Mercer N., (2000), *Words and minds: how we use language to think together*, London: Routledge.

Osborne J., MacPherson A., Patterson A. and Szu E., (2012), Introduction, in Khine M. S. (ed.), *Perspectives on scientific argumentation: theory, practice and research*, Dordrecht: Springer.

Pontecorvo C. and Girardet H., (1993), Arguing and reasoning in understanding historical topics, *Cognition Instruct.*, **11**(3&4), 365–395.

Ryu S. and Sandoval W. A., (2012), Improvements to elementary children's epistemic understanding from sustained argumentation, *Sci. Educ.*, **96**, 448–526.

Sadler T. D. and Fowler S. R., (2006), A threshold model of content knowledge transfer for socio scientific argumentation, *Sci. Educ.*, **90**, 986–1004.

Sampson V., Grooms J. and Walker J. P., (2011), Argument-driven inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: an exploratory study, *Sci. Educ.*, **95**(2), 217–257.

Sampson V., Enderle P. J. and Walker J. P., (2012), The development and validation of the Assessment of Scientific Argumentation in the Classroom, (ASAC) observation protocol: a tool for evaluating how students participate in scientific argumentation, in Khine M. S. (ed.), *Perspectives on scientific argumentation: theory, practice and research*, Dordrecht: Springer.

Sandoval W. A., (2003), Conceptual and epistemic aspects of students' scientific explanations, *J. Learn. Sci.*, **12**(1), 5–51.

Sandoval W. A. and Millwood K. A., (2005), The quality of students' use of evidence in written scientific explanations, *Cognition Instruct.*, **23**(1), 23–55.

Sandoval W. A. and Millwood K. A., (2007), What can argumentation tell us about epistemology? in Erduran S. and Jiménez-Aleixandre M. P. (ed.), *Argumentation in science education: perspectives from classroom-based research*, Dordrecht: Springer.

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

van Eemeren F. H., Garssen B., Krabbe E. W., Henkemans A. F. S., Verheij B. and Wagemans J. M., (2014), *Handbook of argumentation theory: a comprehensive overview of the state of the art*, Springer Academic.

Venville G. J. and Dawson V. M., (2010), The impact of a classroom intervention on grade 10 students' argumentation skills, informal reasoning, and conceptual understanding of science, *J. Res. Sci. Teach.*, **47**(8), 952–977.

Walker J. P., Sampson V., Grooms J. and Zimmerman C., (2011), A performance-based assessment for limiting reactants, *J. Chem. Educ.*, **88**(8), 1243–1246.

Warren B., Ballenger C., Ogonowski M., Rosebery A. S. and Hudicourt-Barnes J., (2001), Rethinking diversity in learning science: the logic of everyday sense-making, *J. Res. Sci. Teach.*, **38**, 529–552.

Wu Y.-T. and Tsai C.-C., (2012), The effects of university students' argumentation on socio-scientific issues via on-line discussion in their informal reasoning regarding this issue, in M. S. Khine (ed.), *Perspectives on scientific argumentation: theory, practice and research*, Dordrecht: Springer.

Yore L. D., Florence M. K., Pearson T. W. and Weaver A. J., (2006), Written discourse in scientific communities: a conversation with two scientists about their views of science, use of language, role of writing in doing science, and compatibility between their epistemic views and language, *Int. J. Sci. Educ.*, **28**, 109–141.

Zembal-Saul C., (2009), Learning to teach elementary school science as argument, *Sci. Educ.*, **93**, 687–719.

