US-China Education Review B, April 2017, Vol. 7, No. 4, 167-181
doi: 10.17265/2161-6248/2017.04.001

A Scientific Inquiry Abilities Test (SIAT) for Senior High School Students

Fu-Pei Hsieh,
Kuang-Hua Primary School, Kaohsiung, Taiwan

Jeng-Fung Hung, Men-An Pan, Kun Lee
National Kaohsiung Normal University, Kaohsiung, Taiwan

The purpose of this study was to develop a reliable and valid Scientific Inquiry Abilities Test (SIAT) for practical and academic uses. Two hundred and thirty 10th, 11th, and 12th grade students in Taiwan provided answers used for item analysis and construct validity testing. The results were as follows: 1. The SIAT contains 27 items including five testlets; 2. The average of difficulty analysis was 0.74 and the average of discrimination analysis was 0.20; 3. The internal consistency was 0.629 and the scorer reliability coefficient was 0.964; 4. There were moderate correlations between all testlet scores and whole test scores (p = 0.000) with values ranging from 0.534 to 0.684; 5. The values for discriminant validity ranged from 0.158 to 0.323; and 6. High-achievement students out-performed their counterparts on the SIAT, and large effect sizes (ESs) were also found for all the testlets. These results indicated that the SIAT possesses high-reliability and construct validity. Finally, the paper addressed implications for future research and science teaching.

Keywords: scientific inquiry, Scientific Inquiry Abilities Test (SIAT), validation

Introduction

The main goal of a science curriculum is to develop students' ability to ask relevant questions and enhance their understanding of inquiry, such that they develop their own scientific thought process, rather than just copying the thoughts and ideas of the scientists they study. This has already become the consensus of the international scientific education field (Abd-El-Khalick et al., 2004). Accordingly, the subject of inquiry is of an urgent concern in science education presently. Hinrichsen and Jarrett (1999) had stressed that students needed to construct, design, conduct, analyze, and communicate their understanding, investigations, and findings. They also need to understand abstract ideas as well as rethink concepts, adapt and retry their own hypotheses, and engage in investigations and problem-solving efforts. Kuhn (2007) had pointed out that argumentation,

Fu-Pei Hsieh, Ph.D., teacher, Kuang-Hua Primary School.
Jeng-Fung Hung, Ph.D., professor, Graduate Institute of Science Education and Environmental Education, National Kaohsiung Normal University.
Men-An Pan, Ph.D. student, Graduate Institute of Science Education and Environmental Education, National Kaohsiung Normal University.
Kun Lee, Ph.D. student, Graduate Institute of Science Education and Environmental Education, National Kaohsiung Normal University.

explanations, model building, and experimentation were the principal approaches to teaching the scientific method. Meanwhile, Koslowski, Marasia, Chelenza, and Dublin (2008) had stressed that in actual scientific inquiry, people sought out explanations for happenings that arose from the covariate. In reality, an explanation becomes increasingly convincing when evidence that connects to that explanation increases. Since inquiry is a way to achieve and understand knowledge about the world, we need a more consistent view for teachers and researchers to follow when exploring related issues. According to the National Research Council (NRC) (1996), basic characteristics of inquiry-based classrooms are:

1. Learners are allowed to invest in scientific problems;
2. Learners are given priority to develop and evaluate explanations and answers to scientific problems using evidence;
3. Learners use evidence to form explanations to answer scientific problems;
4. Learners evaluate their explanations based on other explanations reflected by scientific understanding;
5. Learners exchange and defend their explanations.

Of the above characteristics, Hung (2010) classified 1 and 2 as elements of "inquiry," described 3 as the element of "explanation," called 4 the element of "introspection," and posited 5 as the element of "exchange." Hung also stated that characteristic 4 was related to learners evaluating their own explanations, thinking, and learning, which was related to the regulation of cognition. That is, students should explore, reflect, describe, ask, explain, test, and communicate when they are involved in inquiry.

Literature Review

In the last few years, several articles have been devoted to the study of scientific inquiry (Fradd & Lee, 1999; Ruiz-Primo & Furtak, 2007; Sandoval & Reiser, 2004; Schwarz, 2009). For example, Fradd and Lee (1999) discussed the role of teachers in identifying effective approaches to scientific inquiry and illustrated differences in the ways that teachers perceived and engaged in scientific inquiry with diverse students. Ruiz-Primo and Furtak (2007) presented an eliciting, student response, recognizing, and using information (ESRU) cycle for examining practices based on eliciting, recognizing, and using information for formative assessments and scientific inquiry. They concluded that the ESRU cycle was a useful method for students with higher academic performance. Sandoval and Reiser (2004) identified a framework for scaffolding epistemic aspects of inquiry that can help students understand inquiry processes. Schwarz (2009) cultivated pre-service teachers' principled reasoning through modeling-centered scientific inquiry. The participants were most likely to advance their knowledge and practices through modeling-centered inquiry, provided they were given opportunities to unpack and apply reform-based instructional frameworks. To sum up the above viewpoints and findings, scientific inquiry can enhance students' learning outcomes and knowledge, improve the quality of science teaching, and be of great importance in science education. However, how to examine the effects of students' inquiry-related abilities? This was also an important issue, but none of the above studies really addressed this important consideration. We thought that an appropriate instrument for educators to measure related constructs was being imperative on account of this unaddressed issue, because while those previous studies were fruitful for understanding the teaching of scientific inquiry, they focused largely on collecting qualitative data. In other words, developing a suitable and useful test for assessing students' scientific inquiry abilities quantitatively has always been and continues to be a challenge for educators and teachers.

Previous literature shows that the evaluation process for scientific inquiry-related abilities must consider its content and evaluation method. In terms of the contents of scientific inquiry ability, some scholars believe that, from a scientific process perspective, advocating scientific inquiry ability involves a combination of a set of scientific process skills (Germann, Aram, & Burke, 1996; Keys & Bryan, 2001). Hung (2010) considered that, according to this perspective, students' performance in various scientific process skills should be evaluated. The question design often has nothing to do with the subject-related academic knowledge, different situations, and different scientific process skills are evaluated through different questions. Some scholars, based on the thinking strategies used in scientific inquiry, have advocated that scientific inquiry ability is a type scientific thinking skill (Ben-David & Zohar, 2009; Zion, Michalsky, & Mevarech, 2005). Hung (2010) believed that, according to this perspective, a student's performance in a given inquiry task should be evaluated, and the question design should be relevant to the given subject knowledge and situation. Thus, various skills in scientific inquiry can be evaluated by the same question. However, we believe that scientific inquiry possesses both scientific process skills and scientific thinking skills. When students are conducting scientific inquiry, they must not only utilize relevant scientific procedure skills for problem-solving, but also use previously learned scientific knowledge to conduct scientific thinking. Thus, to test scientific inquiry ability, we must simultaneously test these two types of ability.

In addition, with regards to evaluating scientific inquiry abilities, there are multiple-choice questions, open-ended writing, and actual implementation types (Lawrenz, Huffman, & Welch, 2001). Multiple-choice and open-ended writing questions have lower costs than implementation type evaluations and are suitable for large-scale evaluation (Alonzo & Aschbacher, 2004). Furthermore, in comparisons between multiple-choice and open-ended writing, various scientific inquiry abilities were separated in different situations by multiple-choice test (Wenning, 2007), while open-ended writing can evaluate various scientific inquiry abilities in the same situation (Zion, Michalsky, & Mevarech, 2005). We believe that multiple-choice questions can provide students with limited clues and allow them to select the most appropriate answer from available options. Although writing answers for open-ended questions will be more time consuming, it will allow the students to freely write their own thoughts. This allows the teachers to truly understand students' scientific inquiry abilities. Each type of test has its own value. Therefore, according to the above description, we take into the scientific process account with scientific thinking skills to combine the advantages of multiple-choice and open-ended writing evaluations to develop a reliable and valid Scientific Inquiry Abilities Test (SIAT) for practical and academic uses. Scientific inquiry ability includes the following five elements: (a) identifying patterns and relationships; (b) science process skills; (c) argument and contradiction for competition theory; (d) inquiry and argument; and (e) competition theory and evidence. We used these elements to form the structure of the SIAT. In this study, difficulty analysis, discrimination analysis, discriminant validity analysis, and critical ratio testing were conducted. Furthermore, internal consistency reliability and scorer reliability were also analyzed to validate the SIAT.

Research Design

Subjects

Two hundred and thirty 10th, 11th, and 12th graders were selected from two senior high schools in Kaohsiung in Taiwan to participate in item and construct validity testing. The subjects had learned about various relevant concepts, such as control variables, response variables, combustion, induction, the fact that light travels in a straight line, and so forth, in their science textbooks.

When students are using this evaluation tool for measuring, they need to utilize previously learned scientific knowledge to conduct thinking experiments. Therefore, there should be a high-correlation between their scientific academic achievement and their scientific inquiry ability measurement results. In other words, students with high science academic achievement should have high scores in the SIAT, whereas students with low science academic achievement should have low SIAT scores. That is why the participants are divided into two groups for testing the construct validity of the SIAT. The scores in the top, 25% are the high-achievement group, while those in the lowest, 25% are the low-achievement group. Consequently, there were 70 students identified as high-achievers and 67 students identified as low-achievers on the basis of the SIAT.

Instrument

The SIAT contains 27 items including five testlets. The questions are arranged from simple to difficult and in a hierarchical manner to prevent students from feeling frustrated when answering. Every testlet included a picture or an article about the problem situation, and after reading the directions, the participants answer the questions which are derived from the situation (see Appendix). The outline for the testlet is as follows:

The first testlet is called “amoeba,” which includes two items to test the “identifying patterns and relationships” element (Lawson, 2010). Since the amoeba’s tail is outside the frame, the students must see the relationship between the tail and the frame to select the correct answer. The sub-score was 12. Five amoebas and six non-amoebas were provided by authors, and six pictures of creatures were presented. The students were asked to judge if the pictured creature was an amoeba or non-amoeba. For example, Item one: Look at the picture. Which one is an amoeba? Please explain your reasons.

The second testlet is called “seawater freezing,” which includes nine items to test the “science process skills” element. Eight of the items used were published by Prentice (2000), while the ninth item was designed by the authors. The sub-score was nine. The relevant procedural skills measured include: conclusions, hypotheses, observations, experiments, problem descriptions, data, variable manipulation, and responding variables. Nine explanations were provided and nine questions were asked in reply to these explanations. For example, Item one: According to the explanations, which one is the conclusion?

The third testlet is called “candle burning experiment,” which includes five items to test the “argument and contradiction for competition theory” element. The testlet is based on the aforementioned paper by Lawson (2010). The sub-score was 11. Two related theories were proposed as the explanation for a candle burning and the students were then asked to select the correct theory. Theory A involves scientific knowledge learned by students from textbooks: Combustion will consume the oxygen in the air and make the pressure in the glass smaller. Therefore, the water level in the glass will rise. However, the volume of the rising water level is different from the volume of oxygen consumed in the air. Theory B is the correct explanation: The air around the candlewick expands from the heat of the flame. The air covered by the glass is hot air. When the flame extinguishes, the temperature will drop, reducing the air pressure in the bottle. The atmospheric pressure will then push the water in the bottle. In the testlet, an experimental design and two competitive theories—theory A and theory B, are provided by scientist A and scientist B, and five questions are asked in reply to the explanations. For example, Item two: How can you prove that theory B is wrong? Imagine that you are scientist A.

The fourth testlet is called "salmon returning home," which includes five items to test the "inquiry and argument" element. This testlet uses the narrative method to present itself (Lawson, 2010), allowing students to conduct probabilistic thinking on real experimental data. To verify the relationship between the hypotheses and the evidence, students must select the correct option for the meaning of the subject. However, the question situation is slightly different from in the second question and involves a more advanced topic. The sub-score was eight. In the testlet, one experimental situation and some findings about "salmon returning home" were provided by the authors, and five questions were asked in reply to the findings. For example, Item five: Does the evidence support the biologist's hypothesis? Please explain your reasons.

The fifth testlet is called "competition theory," which includes six items to test the "competition theory and evidence" element. The vignette used is based on a paper by Osborn (2004), and the student misconceptions mentioned in that study were used for descriptions of the questions. For example, if there is no light, we cannot see a thing, use diversified evidence for students to make and explain the reasons. The sub-score was 18. In the testlet, two competing theories which answered the question "How does light move forward?" were presented, along with five pieces of evidence and one conclusion to be judged according to the theories.

> Evidence 2: At night when there is no sunlight, we can still see the object.
> (a) Support Theory A (b) Support Theory B (c) Support Theories A and B (d) Do not support Theories A and B
> Please explain your reasons.

Procedure and Data Analyses

This testing tool's testing time is approximately 50 minutes, accounting for one class. During the testing process, it was discovered that all the students could complete the entire tool within the allotted time. In addition, because the scope of scientific inquiry ability includes many different abilities to allow teachers to use this testing tool with more flexibility, individual sub-tests can also measure the independent ability, so teachers can decide whether to use the whole test or a given testlet. Because the whole test is designed for investigating students' scientific inquiry abilities, the higher the score they get, the stronger scientific inquiry abilities they possess. The worse the score they get, the lower the scientific inquiry abilities they possess.

Difficulty analysis, discrimination analysis, internal consistency reliability, scorer reliability, discriminant validity, and critical ratio were tested for the validation of the SIAT. Furthermore, effect size (ES) is a name given to a family of indices that measure the magnitude of a treatment effect. Unlike significance tests, these indices are independent of a sample. According to Cohen (1988), a *D*-value of 0.2 indicates only a small effect, 0.5 indicates a medium effect, and 0.8 or greater indicates a large effect. This means that if two groups' means do not differ by 0.2 standard deviations or more, the difference is trivial, even if it is statistically significant.

Data from initial responses to each question in the test were accumulated on a spreadsheet to develop the main categories. The two researchers independently performed a micro-analysis on test responses to identify patterns, and then translated these patterns into categories. The researchers then met to discuss differences in categories and grouping criteria. Through discussions, categories were added or discarded and groupings were modified, so that criteria reasonably similar were grouped together under one category. After meetings, consensus was reached and a coding scheme was established that fully satisfied both researchers. Under the appropriate category, a scoring mechanism was used to organize the explanations students provided for their responses.

For example, there are six options for students to select on the first question of testlet one. Each item is worth one point with a full score of six points. If their answer is the correct answer or if they did not pick the wrong answer, then they get one point. The correct answers are 1, 2, and 6, while the others are incorrect. For example, if the student's answers are 1, 2, and 6, then he/she can get six points. If the student's answers are 1, 2, 3, and 6, then he/she does not get a point for answer 3, because it is a wrong answer, meaning the student made an error in judgment. He/she would then get five points. If the student's answers are 2, 3, and 6, then two points will not count, because the student would have made two wrong choices. He/she will get four points. On the second question, there are three reasons for the relevant explanation: Amoebas have a flagellum (tail), spots, and eyes. Writing a correct reason will earn two points. Thus, if students can write all the answers, then they can get six points. If students can write only one answer, then they only get two points.

Findings and Discussion

Coaley (2010) pointed out that if the *p*-value was high, then the item might be too easy. On the other hand, a very low-value indicates that the item may be too difficult. The mean item *p*-value of for a moderate difficulty level is about 0.5. National Education Goals Panel (2004) divided difficulty levels into equal quartiles, labeling the quartiles: "Easy," "Moderate," "Challenging," and "Very challenging." Consequently, the values for "Easy" were from 0.76 to 1, "Moderate" values were from 0.51 to 0.75, "Challenging" values were from 0.26 to 0.50, and "Very challenging" values were from 0 to 0.25. Difficulty and discrimination analysis results for the SIAT are indicated in Table 1. Difficulty analysis meant using the correct ratio ($P = \frac{R}{N} \times 100\%$), while discrimination analysis involved sub-tracting the two difficulty levels ($D = P_H - P_L$). The first question in the first testlet was encoded as Item 101. Accordingly, the second one in the first testlet was Item 102 and the third one in the third testlet was Item 303.

All values of difficulty analysis were from 0.006 to 0.996, and the average difficulty analysis value was 0.74. Difficulty analyses values for 11 questions were higher than 0.9, four questions were higher than 0.8, two questions were higher than 0.7, two questions were higher than 0.6, and three questions were higher than 0.5. In other words, there were 22 questions higher than 0.5 and only Item 206 was lower than 0.1. Because difficulty analyses values for these questions were higher than 0.5, they were easy to respond to for most students, while the value for Item 206 was 0.006, it may be harder for most students. Sixteen items (59%) were labeled as "Easy," six items (22%) as "Moderate," four items (15%) as "Challenging," and only one item (4%) as "Very challenging." This means that more than half of the participants could answer most questions on the SIAT, so whole items were appropriate for all participant ability levels.

All values of discrimination analysis were from 0.002 to 0.565 with the average value being 0.20. The *D*-values of 11 questions were higher than 0.2 and only nine questions were lower than 0.1. Coaley (2010) pointed out that that if the *D*-value was high and positive, then the item might discriminate between high- and low- scorers. In the present study, all the values were positive, revealing that discrimination analysis was suitable for differentiating between high- and low- levels.

Table 1

Difficulty and Discrimination Analysis for the SIAT

Item	101	102	201	202	203	204	205	206	207
P (%)	99.1	91.7	54.3	81.3	32.6	36.1	94.3	6.1	69.6
D (%)	1.5	25.4	13.8	10.9	6.0	12.9	6.1	10.0	35.0
Item	208	209	301	302	303	304	305	401	402
P (%)	53.5	49.1	95.7	82.6	79.6	70.0	44.8	93.0	83.5
D (%)	48.8	56.5	9.0	27.0	40.5	30.7	30.1	6.1	25.6
Item	403	404	405	501	502	503	504	505	506
P (%)	57.4	61.3	87.4	92.2	97.0	96.5	99.6	95.7	97.4
D (%)	34.1	55.8	13.8	19.5	4.6	10.4	1.5	9.0	6.0

Mitchell and Jolley (2009) pointed out that an acceptable internal consistency value needed to be at least 0.7, not involve a measure with a Cronbach's alpha of 0.5, and that the scorer reliability coefficient should be from 0 to 1 with a score of around 0.9 being generally expected. The higher scorer reliability means that the measure will be objective and it puts a ceiling on overall test-retest reliability. In this study, estimates of the internal consistency reliability of the whole test calculated using Cronbach's alpha coefficient, were satisfactory for the samples. The internal consistency (α-reliability) was 0.629 and the scorer reliability coefficient was 0.964. Although the 27 open-ended questions of the SIAT might be expected to impact the participants' replies, the value of Cronbach's alpha was still higher than 0.5, and the scorer reliability was higher than 0.9. In addition, another method used in construct validity testing is checking the extent of internal consistency by correlating specific sub-tests with the test's total score (Groth-Marnat, 2009). There were moderate correlations between all testlet scores and whole test scores ($p = 0.000$), with the values ranging from 0.534 to 0.684 (see Table 2, the data in bold type). This means that homogeneity testing showed moderate values and that the testlet scores and whole test were homogeneous.

Table 2

Discriminate Validities and Homogeneity Test for the SIAT

Testlet	One	Two	Three	Four	Five
One	1				
Two	0.071	1			
Three	0.178**	0.247**	1		
Four	0.265**	0.158*	0.260**	1	
Five	0.168*	0.323**	0.202**	0.209**	1
Total	**0.615****	**0.534****	**0.562****	**0.684****	**0.607****

Notes. $^{*}p < 0.05$; $^{**}p < 0.01$.

In the views of Carducci (2009) and Groth-Marnat (2009), discriminate validity was the degree to which a measure was not too related to another measure. An unrelated measure would show low or negative correlations with variables that were not like it. Although the *α*-reliability coefficients confirmed that there was high internal consistency in this test, it was necessary to determine whether the testlets overlapped. There were low-correlations between testlet pairs (p = 0.000) with the values ranging from 0.158 to 0.323 (see Table 2, the data in normal type). This means that the discriminate validity values are low-values and that the testlets measure distinct.

The critical ratio is an index number for discrimination. With independent-samples *t*-testing, the difference between the mean scores of the upper and lower groups should be significant. That means the items can be identified from the reaction of different subjects (Lin, Guo, & Tu, 2009). Table 3 shows values for the independent-samples t-testing of high- and low- achievement participants with the SIAT questionnaire. On the first testlet, the mean of the high-group was 11.76 and the *SD* was 0.69, while the low-group's mean and *SD* were 8.90 and 3.10, respectively. The two student groups did show a significant difference (t = 7.39, p = 0.000). On the second testlet, the high-group's mean was 5.64 and the *SD* was 1.42, while the low-group's mean and *SD* were 3.64 and 1.44, respectively. This showed a significant difference (t = 8.20, p = 0.000). On the third testlet, the high-group's mean was 6.31 and their *SD* was 1.36, while the the low-group's mean and *SD* were 3.46 and 1.48, respectively. There was a significant difference (t = 11.76, p = 0.000). On the fourth testlet, the high-group's mean was 5.63 and the *SD* was 1.24, while the low-group's mean and *SD* were 3.82 and 1.50, respectively. There was a significant difference (t = 7.71, p = 0.000). On the fifth testlet, the high-group's mean was 11.96 and the *SD* was 1.75, while the low-group's mean and *SD* were 8.18 and 1.98, respectively. This revealed a significant difference (t = 11.86, p = 0.000). The *D*-values were from 1.29 to 2.03. For the whole test, the mean of the high-group was 41.30 and the *SD* was 1.94, while the low-group's mean and *SD* were 28.00 and 4.03, respectively. The student groups labeled by achievement for the SIAT did show a significant difference in their total SIAT scores (t = 24.45, p = 0.000), and the *D*-value was 4.24. This indicated that high-achievement students out-performed their counterparts on SIAT and all testlets with large ESs. That is, SIAT is valid for differentiating between high- and low- achievement levels.

Table 3

Independent Samples T-Test of High- and Low- Achievement Students on the SIAT Questionnaire

Testlet	High-group (*N*= 70)		Low-group (*N* = 67)		*t*	*p*	*d*
	Mean	*SD*	Mean	*SD*			
One	11.76	0.69	8.90	3.10	7.39	0.000	1.29
Two	5.64	1.42	3.64	1.44	8.20	0.000	1.40
Three	6.31	1.36	3.46	1.48	11.76	0.000	2.01
Four	5.63	1.24	3.82	1.50	7.71	0.000	1.32
Five	11.96	1.75	8.18	1.98	11.86	0.000	2.03
Total	**41.30**	**1.94**	**28.00**	**4.03**	**24.45**	**0.000**	**4.24**

The Limitation of the Study

However, one limitation of this study was the lack of measurement of other scientific inquiry abilities, because there are a variety of other scientific inquiry capabilities, such as experimental design abilities

(Hinrichsen & Jarrett, 1999) and model building abilities (Kuhn, 2007). However, this study aims to develop a mass measuring tool that can be used in classrooms and research. Therefore, in a limited testing time situation, not all scientific inquiry ability can be included. Only through the viewpoint of combining the scientific process skills and scientific thinking can certain important inquiry skills be tested, including the following: (a) identifying patterns and relationships; (b) science process skills; (c) argument and contradiction for competition theories; (d) inquiry and argument; and (e) competition theories and evidence.

Another limitation of this study was that although the multiple-choice method is easy to use, it can only provide limited answers for the students to choose from. This cannot cover all the ideas of the students. If students have other ideas, they have to give up their own original thought, because it was not provided in the choices and had to choose from the selection on the test tool. In addition, the study discovered after the testing that although open-ended questions allow students to express their own ideas, but students that are not good at expressing themselves, or are slow in writing, and therefore have difficulties in answering the questions. This resulted in some questions being left blank. So, when this type of situation occurs, the teacher can supplement interviews or implementation method, so the teacher may understand whether the blank answer part is due to not knowing the answer or due to difficulties in self expression. This will allow the science inquiry test results to be more complete.

Conclusions

Recently, science teachers and educators have agreed that inquiry should be taught, because inquiry ability helps students learn more about the world. Since students' inquiry abilities mostly comes from their learning results in science courses, the teacher can test students' inquiry abilities to understand whether their own teaching goals have been reached. This is a chance to use the results to reflect on their own teaching, and thus, improve their teaching methods and strategies. Therefore, investigating students' inquiry ability is not the only reason to develop a SIAT. Doing so will also assist teachers in teaching and educators in their research. It is important to note that creating an instrument for examining students' inquiry ability is both difficult and worthwhile. However, from past-related literature, we found that most methods used to measure science inquiry ability tend to favor performance assessment. This requires more time for testing, but with limited teaching time, inconveniences will be caused for many teachers.

Moreover, most implementing situation can only provide a single question situation for the students to conduct inquiry. However, can this single situation test result be analogous to other problem situations? In other words, will the student show the same inquiry ability in other question situation as he/she did in this question? Therefore, we hope to improve the difficulties and problems of past scientific inquiry ability test in this research. Combining related theories and approaches, we created a quantitative instrument for testing scientific inquiry abilities of students in Taiwan; it helps to examine the extent to which some dimensions of scientific inquiry are consistent from a reform perspective. In this study, NRC's (1996) viewpoint was still used to assess inquiry ability using multiple-choice and open-ended questions. The conclusions are as follows:

First of all, the *p*-values for the 22 questions were higher than 0.5 according to difficulty analyses, and the mean item *p*-value was 0.74. The difficulties of the items were distributed as follows: Fifty-nine percent were labeled as "Easy," 22% as "Moderate," 15% as "Challenging," and only 4% as "Very challenging." In terms of discrimination analyses, the D-values for the 22 questions were from 0.015 to 0.565, and the mean item D-values was 0.20. We found that discrimination analyses were suitable for differentiating between high- and low- levels. Those results revealed that the SIAT has appropriate difficulty and discrimination analyses.

Secondly, the α-reliability coefficient confirmed that there was high internal consistency for the whole test: the α-value was 0.629 and the scorer reliability coefficient was 0.964. In addition, there were moderate correlations between all the testlet scores and whole test scores with the values were from 0.534 to 0.684. The testlets also measured distinct for low-correlations existed between testlet pairs with values ranging from 0.158 to 0.323. This means that he discriminate and convergent validities of the SIAT were also found to be acceptable.

Finally, there were significant differences between the high- and low- achievement groups for the whole SIAT and all the testlets, with large ES indicated for all. The high-achievement students out-performed their counterparts on the SIAT. This revealed that the SIAT is valid for differentiating between high- and low- levels. In summary, these results indicated that factor analysis, internal consistency, and discriminant validity were found to be acceptable, meaning that the SIAT has quality reliability and validity.

References

Abd-El-Khalick, F., BouJaoude, S., Duschl, R. A., Hofstein, A., Lederman, N. G., Mamlok, R., ... Tuan, H. (2004). Inquiry in science education: International perspectives. *Science Education, 88*(3), 397-419.

Alonzo, A. C., & Aschbacher, P. R. (2004, April). Value-added? Long assessment of students' scientific inquiry skill. In *The Annual Meeting of the AERA*, San Diego.

Ben-David, A., & Zohar, A. (2009). Contribution of meta-strategic knowledge to scientific inquiry learning. *International Journal of Science Education, 31*(12), 1657-1682.

Carducci, B. J. (2009). *The psychology of personality: Viewpoints, research, and applications* (2nd ed.). Malden, M.A.: Wiley-Blackwell.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. London, U.K.: SAGE Publications Ltd..

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York, N.Y.: Academic Press.

Fradd, S. H., & Lee, O. (1999). Teachers' roles in promoting science inquiry with students from diverse language backgrounds. *Educational Researcher, 28*(6), 14-20.

Germann, P. J., Aram, R., & Burke, G. (1996). Identifying patterns and relationships among the responses of seventh-grade students to the science process skill of designing experiments. *Journal of Research in Science Teaching, 33*(1), 79-99.

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, N.J.: John Wiley & Sons.

Hinrichsen, J., & Jarrett, D. (1999). *Science inquiry for the classroom: A literature review*. Portland: Northwest Regional Educational Laboratory.

Hung, J. F. (2010). The influences of a thinking-based inquiry learning intervention on eighth graders' scientific inquiry abilities. *Chinese Journal of Science Education, 18*(5), 389-415.

Keys, C. W., & Bryan, L. A. (2001). Co-constructing inquiry-based science with teachers: Essential research for lasting reform. *Jorunal of Research in Sceince Teaching, 38*(6), 631-645.

Koslowski, B., Marasia, J., Chelenza, M., & Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cognitive Development, 23*(4), 472-487.

Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education, 91*(5), 710-726.

Lawrenz, F., Huffman, D., & Welch, W. (2001). The science achievement of various subgroups on alternative assessment formats. *Science Education, 85*(3), 279-290.

Lawson, A. E. (2010). *Teaching inquiry science in middle and secondary schools*. Thousand Oaks, C.A.: SAGE Publications Inc..

Lin, Y. M., Guo, S. J., & Tu, C. A. (2009). The development of inventory for cooperative learning attitude of vocational senior high school students. *International Journal of Technology and Engineering Education, 6*(1), 29-37.

Mitchell, M., & Jolley, J. (2009). *Research design explained* (7th ed.). New York, N.Y.: Harcourt Brace College Publishers.

National Education Goals Panel. (2004). *National education goals report: Building a nation of learners*. Washington, D.C.: Government Printing Office.

National Research Council (NRC). (1996). *The National Science Education Standards*. Washingtion, D.C.: National Academy Press.

Osborne, J., Simon, S., & Erduran, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020.

Prentice, H. (2000). *Integrated science laboratory manual* (teacher's ed.). Englewood Cliffs, N.J.: Prentice-Hall Inc..

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57-84.

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education, 88*(3), 345-372.

Schwarz, C. (2009). Developing preservice elementary teachers' knowledge and practices through modeling-centered scientific inquiry. *Science Education, 93*(4), 720-744.

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. *Journal of Physics Teacher Education Online, 4*(2), 21-24.

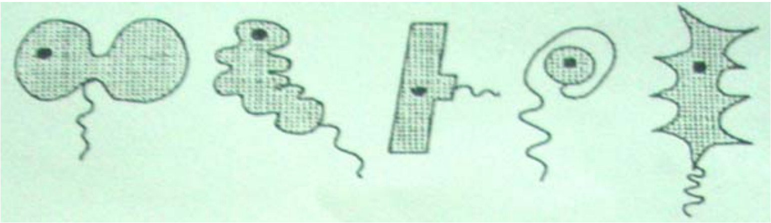
Zion, M., Michalsky, T., & Mevarech, Z. R. (2005). The effects of metacognitive instruction embedded within an asynchronous learning network on scientific inquiry skills. *International Journal of Science Education, 27*(8), 957-983.

Appendix: Items of the SIAT

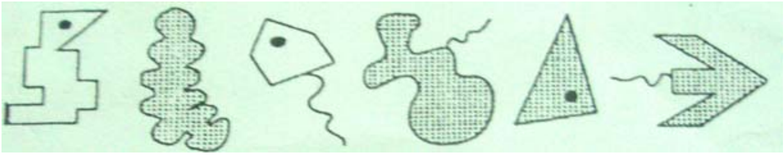
First question: Amoeba

Which of the creatures numbered do you think is/are amoeba(s)? Please explain your reasons.

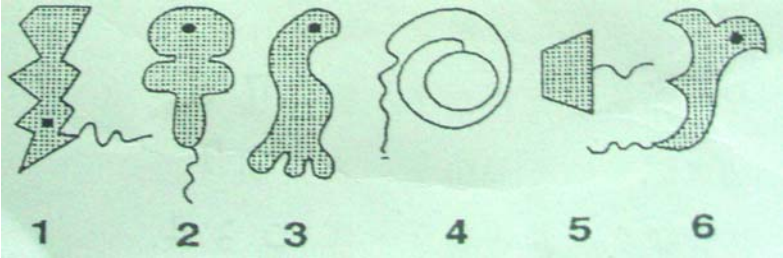
All the below organisms are amoebas.



All the below organisms are not amoebas.



Which of the below organisms are amoebas?



Second question: Seawater freezing

Read the following statements and answer the related questions:

1. A scientist wants to find out why seawater’s freezing temperature is lower than that of freshwater.

2. The scientist entered the library and read information about the physical properties of the liquid.

The scientist also read articles associated with seawater.

The scientist went to the beach, observed the environmental conditions, and took notes of the taste of the seawater and the surrounding environmental factors, such as wind, waves, pressure, temperature, and humidity.

After collecting the above information and data, the scientist made this assumption: Because seawater contains salt, it has a lower freezing temperature than freshwater.

The scientist went back to the lab and conducted the following experiment:

In two separate beakers, the scientist injected 1L of freshwater.

He put 35 g of salt in one of the beakers.

The scientist placed these two beakers together in a -1 °C environment for 24 hours.

After 24 hours, the scientist checked the two beakers and found that the freshwater has frozen, but the saltwater is still liquid.

The scientist wrote in the notebook: This phenomenon shows that the saltwater has a lower freezing temperature than freshwater.

The scientist continued to write: The freezing point of seawater is lower than the freezing point of freshwater, because seawater contains salt and freshwater does not.

After each question, please answer the question with the above described numbering:

1. What conclusions were included in the above statements? Answer:________________________

2. What assumptions were included in the above statements? Answer:________________________

3. What observations were made in the above statements? Answer:__________________________

4. How many experiments were conducted in the above statements? Answer:__________________

5. In which of the above statements was the question posed? Answer:________________________

6. Which of the above statements contain data? Answer:___________________________________

7. What was the independent variable in this experiment? Answer:___________________________

8. What was the dependent variable in this experiment? Answer:____________________________

9. What was the control variable in this experiment? Answer:_______________________________

Third question: Candle burning experiment

As shown in the following figure, a candle is fixed on the bottom of a sink. The sink is filled with 2 cm of water and the candle is lit. A glass bottle is then used to cover the candle. The candle flame gradually became smaller until it is extinguished. At the same time, the water in the bottle rose. Scientist A and B each proposed a different theory to explain this test result:

Theory A: The oxygen in the glass bottle as consumed by combustion so the candle flame went out. The oxygen in the air of the bottle was used up, causing the water in the glass bottle to rise.

Theory B: The air near the wick is heated by the flame and the glass bottle covered the hot air. After the flame went out, the temperature dropped. The air pressure in the bottle became smaller and the air pressure pressured the water into the bottle.

Please answer the following questions:

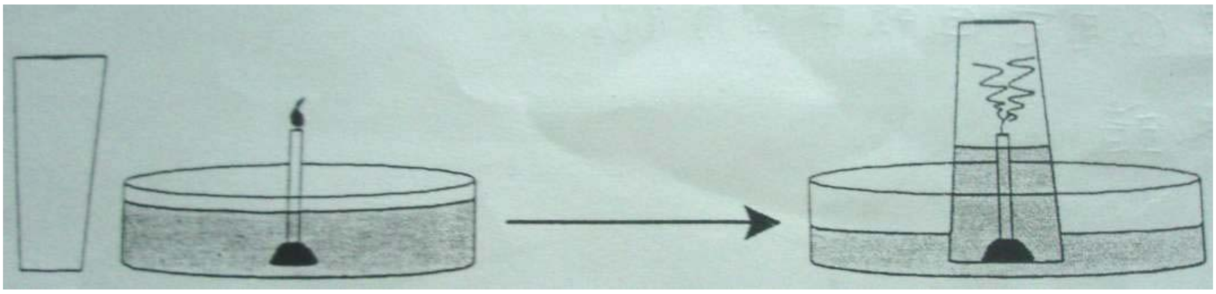
1. If you are scientist A, what evidence do you have to support that theory A is correct?

2. If you are scientist A, what evidence do you have to point out that theory B is incorrect?

3. If you are scientist B, what evidence do you have to support theory B is correct?

4. If you are scientist B, what evidence do you have to point out that theory A is incorrect?

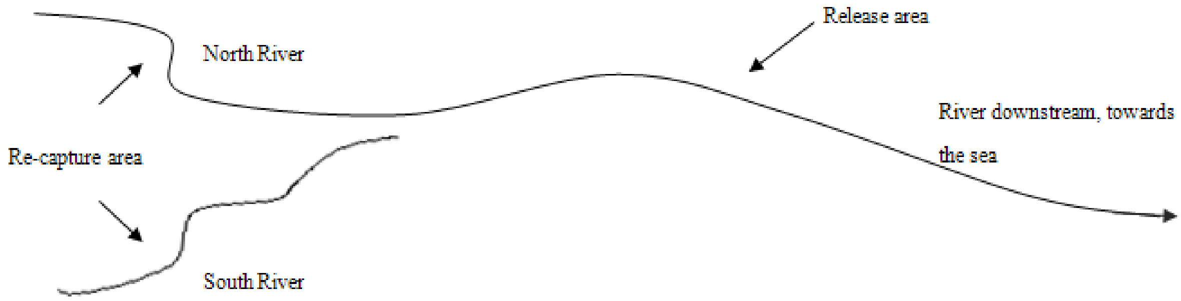
5. According to your answers in the above-mentioned problems, do you support theory A or B theory? Please explain your reasons.



Fourth question: Salmon returning home

Every year during a fixed season, mature salmon will gather from around the ocean and scramble upstream towards their home stream. The goal is to return to the place of their birth and repeat the circle of life. What methods does salmon use to guide themselves from the ocean back to their home river source? A biologist conducted the following experiment to understand this phenomenon. He first caught salmon in North River and South River, and then covered the nose of these salmon to be used as the experimental group. The rest of the salmon captured from North and South River did not have their nose covered. Using these salmon as the control group, he then marked these two groups of salmon for distinction, and then took the salmon downstream to the convergence point of the two rivers and released them. These salmon will swim upstream to their birthplace in North or South River. The data of the re-captured salmon by the biologist in North and South River is as follows:

1. What is the assumption that this biologist is testing?
2. What is the independent variable in this biologist's experiment?
3. What is the dependent variable in this biologist's experiment?
4. What is the control variable in this biologist's experiment?
5. Does the data from the above experiment support his assumptions? Please explain your reasons.



Control group

Re-capture area after release

Capture area	South River	North River
South River	22	3
North River	8	23

Experimental group

Re-capture area after release

Capture area	South River	North River
South River	12	13
North River	11	14

Fifth question: Competition theory

Theory A: Light travels from our eyes to the object, so that we are able to see the object.

Theory B: Light is generated by a light source and reflected from an object to our eyes, so we are able to see the object.

There are five evidence below. Please determine whether each individual evidence supports theory A, theory B, support both, or do not support either:

Evidence 1: Light travels in a straight line.

(a) Support Theory A (b) Support Theory B (c) Support Theories A and B (d) Do not support Theories A and B

Please explain your reasons.

Evidence 2: At night when there is no sunlight, we can still see the object.

(a) Support Theory A (b) Support Theory B (c) Support Theories A and B (d) Do not support Theories A and B

Please explain your reasons.

Evidence 3: Wearing sunglasses can protect our eyes.

(a) Support Theory A (b) Support Theory B (c) Support Theories A and B (d) Do not support Theories A and B

Please explain your reasons.

Evidence 4: If there is no light, we cannot see objects.

(a) Support Theory A (b) Support Theory B (c) Support Theories A and B (d) Do not support Theories A and B

Please explain your reasons.

Evidence 5: You have watched the object in order to see it.

(a) Support Theory A (b) Support Theory B (c) Support Theories A and B (d) Do not support Theories A and B

Please explain your reasons.

Based on your judgments of evidences 1 to 5 in the above, your conclusion is:

(a) Support Theory A (b) Support Theory B (c) Support Theories A and B (d) Do not support Theories A and B

Please explain your reasons.