Jour. Korean Earth Science Society, v. 24, no. 2, p. 93–99, February 2003

Development of the Test Instrument to Assess Students' Progress in Understanding Nature of Science: Based on AAAS Benchmarks for Science Literacy

Eun Ah Lee* and Seung-Urn Choe

Department of Earth Science Education, Seoul National University

Abstract : The purpose of this study was to develop a new test instrument based on AAAS Benchmarks for Science Literacy, to assess k-12 students' progress in understanding nature of science (NOS). A total of 276 items were developed including 33 items for grade k-2, 36 items for grade 3-5, 78 items for grade 6-8 and 129 items for grade 9-12 and they were reviewed for validity and reliability. Key ideas that were the foundation of test items were extended, sophisticated and enriched according to the grade level. The general score of this test represents a student's cognitive state about an understanding of NOS. The result of this test can be expected to give some useful information for follow-up investigations, improving instructional design, and conducting further studies.

Keywords : nature of science (NOS), assessment of NOS conceptions, NOS test instrument

Introduction

The purpose of this study was to develop a new test instrument based on American Association of Advancement for Science (AAAS, 1993) Benchmarks for Science Literacy, to assess k-12 students' progress in understanding nature of science (NOS).

NOS has long been considered as one of the important goals in science education. In spite of such a long history and strong emphasis, however, the definition of NOS is still in hot debate. And the discord about the definition of NOS resulted in ambiguity and confusion among test instruments used in many previous studies.

The one interesting fact is that the result of the research concerning students' and teachers' conceptions of NOS has been consistent. In any result, students and teachers were shown to have an inadequate understanding of NOS (Lederman, 1992).

It would seem improbable that research conclusions would be so consistent if a specific weakness from each instrument was significant (Lederman *et al.*, 1998). It suggests that all of those instruments share something in common at a certain degree. In other words, there is a general agreement about NOS although there are various views.

The new test instrument contains a general consensus of NOS, which is relevant for pre-college students (Akerson *et al.*, 2000), because it was based on the Benchmarks, one of the renowned science education standards documents. The Benchmarks contains a general agreement about NOS. As seen above, the differences among various views of NOS do not significantly affect the standards for k-12th students.

And the new test instrument focused on assessing students' progress in understanding NOS. Most of the previous test instruments were designed for a specific age group, usually adults or high school students. It was not possible to see the progress of students according to their ages. Driver *et al.* (1996) studied the conception of children under 16 years old and showed that the children's conception of NOS changes as they grow. Older children were likely to have a more sophisticated view of NOS. The Benchmarks focuses on students' progress toward science literacy including NOS and provides statements indicating what students

*Corresponding author: eunahj@hanmail.net

should know at the end of a certain grade. How the ideas in the Benchmarks develop according to grades is well shown in Atlas of Science Literacy (AAAS, 2001).

This study was conducted to develop the new test instrument designed to assess the progress of students' NOS conceptions, therefore it was based on the Benchmarks statements which indicate the threshold of NOS conceptions according to grades.

Development of the New Test Instrument

Fig. 1 shows the procedure of the developing the new test instrument and provides overall view of the whole procedure from setting the purpose to reviewing items.

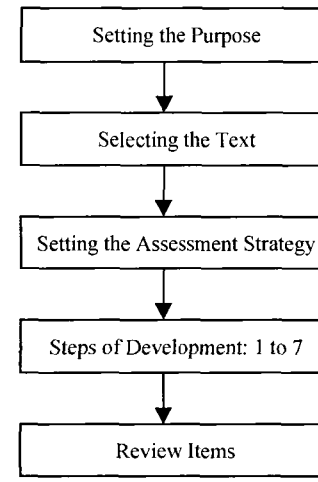

Fig. 1. The procedure of developing the new test instrument

The new test instrument was trying to reduce the ambiguity between the perceptions of students and the perceptions of researchers. Students do not perceive and interpret the test statements in the same way as researchers do (Akikenhead & Ryan, 1992). To avoid ambiguity, the new test instrument adopted three different types for each item, and these types were devised from the Benchmarks. In addition to presenting what students should know at the end of a certain grade, the Benchmarks states that "know" implies:

-Students can explain the idea in their own words.

-Students can relate the idea to ideas in other parts of Benchmarks.

-Students can apply the idea in the novel context (AAAS, 1993).

Based on this implication, the new test items were developed in three types: E-type, R-type and A-type.

An E-type item asks a student to **explain** the idea. An R-type item asks a student to **relate** the idea to other key ideas in the Benchmarks, and an A-type item asks a student to **apply** the idea in new context.

Here is an example. Following three items are developed from the same key idea and each of them asks a student to explain, to relate, and to apply the key idea.

E-type: Scientists assume that the universe is a vast single system in which the basic rules are the same everywhere.

R-type: The stars differ from each other in size, temperature, and age, but they appear to be made up of the same elements that are found on the earth and to behave according to the same physical principles.

A-type: Knowledge gained from studying one part of the universe is applicable to other parts

A student who understands the key idea is supposed to answer all three types of item with consistency. These three types of item might not help respondents to understand the statements better, however, they will help researchers analyze the results. For example, a students who gives an acceptable answer to E-type item but fails in R-type and A-type cannot be considered to know that idea completely.

The test instrument took the form of an item pool and adopted a true-false item format. As Laugksch and Spargo (1996) mentioned, it could be currently unfashionable to advocate the use of true-false item format and truly, there is undeni-

able limitation of this format (Roid & Haladyna, 1982; Ebel & Frisbie, 1991).

Nevertheless, true-false item format is considered being occasionally appropriate if properly constructed and reviewed (Ebel, 1972) and many efforts trying to overcome disadvantages of this format were done through the steps of the development.

All of the test items belong to one of three subgroups: the scientific world view, scientific inquiry, scientific enterprise. These three sub-groups were adopted from the Benchmarks sections.

How to set the grade span was a big issue and it seemed natural to accept grade spans in the Benchmarks since these new test items were derived from the Benchmarks statements. But these grade spans were used only for approximate checkpoints and the boundaries between grades were not strictly fixed.

All test items were developed through the following seven steps.

Step 1. A sentence containing a key idea was selected from the Benchmarks.

Step 2. A selected key sentence was transformed to a true form item.

Step 3. A false form item was developed.

Step 4. An idea related to the key idea was chosen.

Step 5. An R-type true-false item was developed.

Step 6. A new sentence containing the key idea was produced.

Step 7. A new sentence was transformed to a true-false item.

Each test item has a serial number that contains useful information about the character of that item. The serial number consists of seven elements including four letters and three digits. The first letter means grade span, H for grade 9-12, M for grade 6-8, E for grade 3-5, and P for grade k-2. The second letter represents sub-groups, A for the scientific world view, B for scientific inquiry, and C for scientific enterprise. The third letter indicates the type of item. The first digit indicates the paragraph of statements in the Benchmarks and the next two digits represent the item number in that paragraph. The last letter says whether the statement is true or false.

For example, HAE101T means that this item is an E-type, true form item for grade 9-12, belongs to the scientific world view section, and the key idea was extracted from the first paragraph in the Benchmarks, chapter 1, section A.

The serial number will give the test designer essential information to select items from the pool. A few examples of test items are presented in Appendix.

A total of 276 items were developed including 33 items for grade k-2, 36 items for grade 3-5, 78 items for grade 6-8 and 129 items for grade 9-12.

Key ideas that were the foundation of test items were extended, sophisticated and enriched according to the grade level. For higher grade levels, test items were sophisticated both in language and in style, while they were respectively simple and basic for lower grade level. To adopt the proper language for each level, several groups of students from different grade levels participated in reviewing and re-writing items.

The respondent of the new test instrument is asked to answer each item in one of following responses; true, false, I don't know, or I don't understand.

Review of the New Test Items

The test items were reviewed for validity and reliability.

The reliability was tested using test-retest method and split-halves method. Each set of items, targeting grade k-2, grade 3-5, grade 6-8 and grade 9-12 was tested independently. 19 elementary school students, 24 middle school students, 28 high school students, and 21 undergraduate students participated in the pilot test. The reliability coefficients were 0.75 for grade k-2, 0.68 for grade 3-5, 0.60 for grade 6-8, and 0.78 for grade 9-12. Even though there is a possibility of overestimating or

underestimating (Bohrnsteadt, 1985), correlation coefficient 0.60 to 0.78 could tell us that it is fairly correlated (Sung, 1995). Thus, it can be said that this new test instrument is modestly reliable.

To see whether the test instrument is assessing what it intends to assess, content validity, construct validity and item validity were checked.

Content validity focuses upon the extent to which the content of an indicator corresponds to the content of the theoretical concept it is designed to measure (Zeller, 1994). Since all test items are faithfully base upon statements of the Benchmarks, the credit of the Benchmarks buttressed the content validity of the new test instrument.

Construct validity focuses on the assessment of whether a particular measure relates to other measures that are consistent with it in a theoretically anticipated way (Zeller, 1994). Two different approaches were made for construct validation.

First, correlation among the three types of item was examined to see the inner structure of the test instrument. It was done by following Zeller (1994)' s three recommendable steps of construct validation and resulted in modestly correlated coefficients (0.46 to 0.78).

The second approach was made to estimate concurrent validity, for concurrent validity gives important information to construct validity (Song, 1997). Independent corroboration of the validity of the specific key ideas in NOS was examined among the Benchmarks, Science for All Americans (AAAS, 1989), National Science Education Standards (NRC, 1996) and the consistency among them also supported the construct validity of the test instrument.

A panel of judges that consisted of a university professor, graduate students, and teachers from middle schools and high schools both in USA and in Korea was asked to review the item validity. The opinions of judges were discussed and items proved irrelevant were revised.

Discussion

This test instrument is not competent-based. Since there is no single definition of NOS, it is meaningless to score the result in that way.

Thus, the general score of this test only represents a student's cognitive state about an understanding of NOS. The student who has a higher score in this test cannot be said to have a better understanding about NOS. Instead, higher score means that he/she has a closer conception about NOS to which the Benchmarks indicates.

In addition to the general score, three sub-scale scores can be obtained: Scientific World View, Scientific Inquiry, Scientific Enterprise. These subscale scores will help to gain more specific information about students' understanding of NOS.

A neutral answer consists of two types: "I don't know" and "I don't understand".

An "I don't understand" answer from respondents would give some information to improve the item both in language and in validity, while an "I don't know" answer would give information to improve instructional design.

To see the progress in students' understanding of NOS, four sets of the test, each of them addressing a different grade, can be given to the different grade level students. The recommended checkpoints for the test are the end of the 2nd grade, 5th grade, 8th grade, and 12th grade (AAAS, 1993).

The test can be designed in various ways according to its purpose.

Interpreting the result is another challenge. Table 1 and Table 2 show an example of interpreting the result. Ten questions in Table 1 were from the test paper used in the pilot test to estimate the reliability. They require the understanding of difference that can happen in the same kind of science investigation. Table 2 shows how each student group answered and how these results could be interpreted.

Table 1. Items addressing difference in the science investigation

1. When people give different reports of the same thing, it is usually a good idea to make some fresh observations instead of just arguing about who is right. _____
2. People may have different explanations of the same thing, so it is important to find whose idea is the best. Other ideas must be dropped._____
3. When a science investigation (observation, experiment, and others) is done the way it was done before, we expect to get exactly the same result._____
4. Results of similar scientific investigations rarely turn out exactly the same._____
5. Measurements are always likely to give slightly different numbers, even if what is being measured stays the same._____
6. When we weigh something repeatedly, we should obtain exactly the same number each time._____
7. Results of similar scientific investigations rarely turn out exactly the same because of flaws in the methods used and mistakes in observations It is always easy to tell what causes differences._____
8. When similar investigations give different results, the scientific challenge is to remove all the differences to get exactly the same results._____
9. Different explanations can often be given for the same evidence, and it is not always possible to tell which one is correct._____
10. Hypotheses are valuable, even if they turn out not to be true, if they lead to fruitful investigations._____

Table 2. The example of interpreting the result

Grade	Item No.	No. of correct answers	No. of total answers	Interpretation
k-2	1	17	19	Confused about the difference that occurs in science investigations
	2	9	19	
	3	10	19	
3-5	4	20	24	Understand the difference that occurs in science investigations in some degree, but believe its cause is "simple mistake" and is detectable
	5	19	24	
	6	20	24	
	7	10	24	
6-8	8	25	28	Understand the difference that occurs in science investigations, but part of them cannot relate the idea to "hypotheses"
	9	22	28	
	10	15	28	

For instance, most of young elementary students answered, "To make some fresh observations is better when people give different reports of the same thing." But almost half of them answered, "It is important to find one right idea among different explanations of the same thing and to drop others." and "The results of the same science investigation always turn out exactly the same." Therefore, it implies that students had some vague conception about the difference in science investigations and it was not firmly based.

Table 2 also shows the change of students' view about the difference occurring in science investigations according to their age. It becomes more organized and sophisticated as they grow.

Conclusion

Progress in science education can be compared to a long journey toward the goal. If understanding NOS is the goal which students should achieve, how they journey toward that goal is important and must be recognized and guided. AAAS sets such a goal by forming Science for All Americans and again, sets the desirable states of progress by forming Benchmarks for Science Literacy.

This instrument was developed to monitor such progresses in students' understanding NOS conceptions. Thus, the most expected outcome of this instrument is to fulfill AAAS purpose in Benchmarks. In spite of the basic limitation of a paper

and pencil test, the result of this test can be expected to give some useful information for follow-up investigations, improving the instructional design, and conducting further studies. As the combined method that takes both quantitative and qualitative style is frequently used in contemporary assessments, the outcome of this instrument could be used as useful pre-data to a follow-up investigation that is likely to be the qualitative study.

The usefulness of the new test instrument in this study needs to be exposed to critique. This instrument is developed and validated to assess students' understanding of NOS. Only actual usage of the instrument would answer the question of whether it fulfills its purpose or not.

More likely, persistent further studies are expected to improve the instrument.

References

Aikenhead, G. S. and Ryan, A. G., 1992, The development of a new instrument: Views on Science-Technology-Society. Science Education, 76, 477-491.

Akerson, V., Abd-El-Khalick, F. and Lederman, N. G., 2000, Influence or a reflective explicit activity approach on elementary teachers' conception of nature of science. Journal of Research in Science Teaching, 37, 295-317.

American Association for Advancement of Science., 1989, Science for all Americans. American Association for Advancement of Science, Washington, DC, 246 p.

American Association for Advancement of Science., 1993, Benchmarks for Science Literacy. Oxford University Press, New York, 448 p.

American Association for Advancement of Science., 2001, Atlas of Science Literacy. American Association for Advancement of Science, Washington, DC, 49 p.

Bohrnsteadt, G. W., 1985, Measurement. In J. Wright and P. Rossi (Eds.), Handbook of Survey Research. Academic Press, New York, 755 p.

Driver, R., Leach, J., Miller, and Scott, P., 1996, Young people's images of science. Open University, Bristol, PA, 172 p.

Ebel, R., 1972, Essentials of educational measurement. Prentice-Hall, New Jersey, 622 p.

Ebel, R. and Frisbie, D. A., 1991, Essentials of educational measurement (5th ed.). Prentice-Hall, New Jersey, 622 p.

Laugksch, R. C. and Spargo, P. E., 1996, Development of a pool of scientific literacy test-items: based on selected AAAS literacy goals. Science Education, 80, 121-143.

Lederman, N. G., 1992, Students' and teachers' conceptions of the nature of science: A review of the research. Journal of Research in Science Teaching, 29, 331-359.

Lederman, N. G., Wade, P. and Bell, R. L., 1998, Assessing understanding of the nature of science: A historical perspective. In W. F. McComas (Ed.), The nature of science in science education: Rationales and strategies. Netherlands: Kluwer Academic Publishers, p. 331.

National Reasearch Council., 1996, National Science Education Standards. National Academy Press, Washington, DC, 272 p.

Roid, G. H. and Haladyna, T. M., 1982, A technology for test-item writing. Academic Press, New York, 264 p.

Song, I. S., 1997, Research methodology. Sangjisa, Seoul, 438 p.

Sung, T. J., 1995, Understanding and application of modern statistics. Yangseowon, Seoul, 508 p.

Zeller, R. A., 1994, Validity. In T. Husen and T. N. Postlethwaite (Eds.), The international encyclopedia of education 11 (2nd ed). Oxford: Pergamon Press. p. 6569.

Appendix

PAE101T When a science investigation(observation, experiment, and others) is done the way it was done before, we expect to get a very similar result.

PAE101F When a science investigation(observation, experiment, and others) is done the way it was done before, we expect to get exactly the same result.

PAR101T When trying to build something, it usually helps to ask someone who has done it before for suggestions.

PAR101F When trying to build something, it always guarantees success to follow the way of someone who has done it before.

PAA101T When we do an observation or an experiment in the way we did it before, we would have a very similar result.

PAA101F When we do an observation or an experiment in the way we did it before, we would have the exactly same result.

EBE101T Scientific investigations may take many different forms, including observing what things are like or what is happening somewhere, collecting samples for analysis, and doing experiments.

EBE101F Scientific investigations take only three different

forms: observing what things are like or what is happening somewhere, collecting samples for analysis, and doing experiments.

HCA101T The historical and cultural roots of the concepts in science, mathematics and technology can be found in early Egyptian, Greek, Chinese, Hindu and Arabic cultures.

HCA101F The historical and cultural roots of the concepts in science, mathematics and technology can be found mostly in early Greek cultures.

Manuscript received December 28, 2002
Revised manuscript received February 17, 2003
Manuscript accepted February 20, 2003