SHU-NU CHANG and MEI-HUNG CHIU*

THE DEVELOPMENT OF AUTHENTIC ASSESSMENTS TO INVESTIGATE NINTH GRADERS' SCIENTIFIC LITERACY: IN THE CASE OF SCIENTIFIC COGNITION CONCERNING THE CONCEPTS OF CHEMISTRY AND PHYSICS

ABSTRACT. Scientific literacy and authenticity have gained a lot of attention in the past few decades worldwide. The goal of the study was to develop various authentic assessments to investigate students' scientific literacy for corresponding to the new curriculum reform of Taiwan in 1997. In the process, whether ninth graders were able to apply school knowledge in real-life problems was also investigated. Over the course of our two-year study, we developed authentic assessments to investigate a stratified random sampling of 1,503 ninth graders' levels of scientific literacy, including scientific cognition, process skills, application of science, habits of mind, nature of science, and attitude towards science. The purpose of this article is to discuss three different formats of authentic assessments: multiple-choice, open-ended, and hands-on test items, which we developed to investigate scientific cognition. To validate the three formats of authentic assessments, students' performance on these three assessments were compared with the science section of Taiwan's Academic Attainment Testing (STAAT), and the values of Pearson correlation coefficient were all at the significant level, ranging from 0.205 to 0.660 ($p < 0.01$). We found that our three authentic assessments were better in evaluating students' authentic abilities in science than standardized tests (such as STAAT). Further authentic assessments, particularly the hands-on activity, benefited low-achieving students. Concerning the common themes tested in the authentic assessments, students performed better in a multiple-choice test than an open-ended test on electricity and heat and temperature. In addition, two themes of chemical reactions and reactions of acid and base with indicators were performed best in a hands-on test than in the other two tests. In this article, we provide evidence that authentic assessments could be developed in different formats to investigate students' scientific cognition as part of the national test. Of these formats, the multiple-choice, open-ended, and hands-on test items are all shown to be sensitive in their evaluation of students' cognition in science.

KEY WORDS: scientific literacy, scientific cognition and authentic assessment

In this new century, many countries not only focus on research in science and technology, but also dedicate time and resources to the improvement of science education. In fact, education reform has become a continuous movement in many countries (Bell, Abd-El-Khalick, Lederman, McComas & Matthews, 2001; Education Act 1996, 1996; McComas, Almazroa & Clough, 1998), and Taiwan is no exception. The curriculum in Taiwan has been revised many times to date. In the past, while the elementary curriculum was integrated, the junior high school curriculum was divided into

* Author for correspondence.

International Journal of Science and Mathematics Education (2005) 3: 117–140

different subjects (e.g., physical science, biology, and earth science). This arrangement made the transition for students from elementary to junior high school difficult. To facilitate this transition, Taiwan began the latest curriculum reform in 1997, and since 2001, has carried out the reform in grades 1–9.

During this period, curriculum guidelines were developed instead of standards, and curriculum integration was promoted, while school-based curriculum development, team teaching, decreasing the time of each lesson, and simplifying the learning content were discouraged. At the same time, the curriculum reform emphasizes the importance of linking the content of the curriculum with life-related context. It is exactly the spirit of scientific literacy. In Miller's opinion, scientific literacy in today's scientific and technological society should consist of the understanding of the norms and methods of science, key scientific terms and concepts, and the impact of science and technology on society (Miller, 1983). Therefore, under the supervision of the Ministry of Education (MOE), curriculum developers worked within several main themes of scientific literacy – including scientific cognition, process skills, nature of science, attitude towards science, habits of mind, and application of science (MOE, 1998) – to guide the development of science curricula.

Cultivating and raising students' scientific literacy in the different disciplines has become the goal of science education worldwide (AAAS, 1993; MOE, 1998), especially since the 1950s. The term of scientific literacy was first used in the late 1950s, for example Paul Hurd used it in a publication entitled "Scientific Literacy: Its Meaning for American Schools." (Laugksch, 2000). It has always been taken for granted that literacy in science could improve individual's lives in a science- and technology-dominated society, and that it can ultimately enhance international competition (Aikenhead & Ryan, 1992; Laugksch, 2000; Thomas & Durant, 1987). According to Laugksch's (2000) view, the importance of scientific literacy has a macroscopic perspective: (1) scientific literacy could improve the economic well-being of a nation in order to promote international competition; (2) when citizens possess an appropriate level of scientific literacy that will sustain the supply of scientists, engineers, and technically trained personnel. Also the promotion of scientific literacy could contribute to the intellectual culture of a society itself (Shortland, 1988). In contrast with a macroscopic view, scientific literacy could enhance individuals' lives as well. Thomas & Durant (1987) think scientific literacy could facilitate any individual's life in a science- and technology-dominated society, such as helping to make personal decisions concerning diet, smoking, vaccination, screening programmers or safety in the home and at work. To

achieve the goal of scientific literacy, teachers and science educators must concentrate on the design of instruction for student learning or on the evaluation of student performance (Aikenhead & Ryan, 1992; Champagne & Newell, 1992; Jenkins, 1992; Laugksch, 2000; Laugksch & Spargo, 1996).

How do we know whether students have developed scientific literacy? Some researchers have developed multiple-choice test items to evaluate 11th and 12th graders' scientific literacy regarding STS (science-technology-society) issues (Aikenhead & Ryan, 1992), and true-false test items for assessing the nature of science, scientific cognition (earth, physical/chemical, life and health sciences), and the impact of science and technology on society (Laugksch & Spargo, 1996). According to Champagne & Newell (1992), American students perform poorly on these types of conventional assessments; therefore authentic tasks should be developed to bring motivation and skills to students as well as to bridge the gap between elementary and junior high school.

Development of Authentic Assessments

Putting scientific concepts into authentic contexts is an important area of science education (Champagne & Newell, 1992; Yerrick, 2000). Before the first reference to "authentic assessment" by Wiggins (1989), the term "authentic" has been formally used in 1988, concerning the context of learning and assessment (Archbald & Newmann, 1988). After that, "authenticity" represented the appropriate, meaningful, significant, and worthwhile forms of human accomplishment (Newmann & Archbald, 1992). In the meantime, according to Newmann and Archbald's idea, authentic achievement ought to involve disciplined enquiry, higher-order thinking and problem solving capacities useful both to individuals and to society, and transfer the mastery gained from school to life afterwards. Likewise this can be seen through Taiwan's educational reform goal and the spirit of scientific literacy (Miller, 1983), which are both intended to cultivate students' abilities such that they can apply them to the real world beyond school.

However, it is difficult to examine whether or not that goal has been achieved through standardized testing. Therefore, for the two-year project described in this article, we developed six authentic assessments based on the ideas of the latest curriculum reforms that have been carried out in grades 1–9 in 2001 to investigate whether ninth graders are able to apply school knowledge in a real-life context (Chiu, 2002). In the past decade, standardized testing and authentic assessment have been perceived as two different cultures (Dori, 2003). Standardized testing usually was more quantitative, and authentic assessment was considered qualitative and al-

ternative (Wolf, Bixby, Glenn & Gardner, 1991). Nowadays, standardized and authentic testing have started to merge for evaluating students' higher-order thinking skills, which teachers have neglected to develop under test-oriented teaching (Dori, 2003). Nevertheless, the national entrance examinations of senior high schools and universities in Taiwan are also a kind of standardized testing. Taiwan's Academic Attainment Testing (TAAT) is the existing national entrance examination of senior high school, which includes the five subjects of science, English, mathematics, society and Chinese. In addition, the test items of TAAT are all multiple-choice, due to the time and cost limitation. There is a total of 60 minutes for students to answer each subject of English, mathematics and Chinese, and 70 minutes for answering each of society and science, ideally each item taking one minute to complete. Although authentic assessment is the trend in Taiwan as well, seems it is still rare and hard to embed scientific concepts in authentic phenomena regarding the science section of TAAT (STAAT), and that was why we decided to develop authentic assessments and bring them into practice as a national examination.

Furthermore, we found that the authentic assessments we developed in the two-year project coincided with Cumming and Maxwell's theory regarding authentic assessment. To evaluate students' learning in the everyday context, Cumming & Maxwell (1999) suggested four types of assessments related to authentic assessment but which show an integrated view of authentic assessment. The intrinsic and latent components of these four assessments are (1) performance assessment, which assesses the ability to use a disciplined-inquiry integration of knowledge to realistic tasks, (2) situated assessment, which assesses a kind of context specific performance, (3) complexity of expertise and problem-based assessment, which assesses the ability to transfer disciplined-inquiry prior knowledge to life beyond school, and (4) competence-based assessment, which assesses the values one has concerning the perspective of vocations (Cumming & Maxwell, 1999). Except for situated assessment, the other three types of authentic assessments fitted well with our six authentic assessments. Performance assessment conformed to our assessment of process skills based on the definition that performance assessment is the execution of some task or process that should be assessed through actual demonstration (Wiggins, 1993). In addition, the theory of complexity of expertise and problem-based assessment is to develop higher-order thinking and problem solving, and an expectation that students can transfer to life what they have learnt in school (Cumming & Maxwell, 1999). Accordingly, the foundation of our assessments of scientific cognition, application of science, and habits of mind tallied with the complexity of expertise and problem-based assessment.

TABLE I

The interpretations of our assessment-related authenticity

Authentic features	Six main themes	Formats
Complexity of expertise and problem-based assessment	• Scientific cognition	Multiple-choice items Open-ended questions Hands-on activity
	• Application of science	Multiple-choice items Open-ended questions
	• Habits of mind	Open-ended questions
Performance assessment	• Process skills	Hands-on activity Multiple-choice items
Competence-based assessment	• Nature of science • Attitude towards science	Open-ended questions Likert scale items

The latent component of competence-based assessment focuses on personal values, which tie to the real world (Cumming & Maxwell, 1999) and our assessments of the nature of science, and attitude towards science, also fitted in with it. Table I shows the three categories of authentic assessments linked to our authentic assessments for evaluating ninth graders' scientific literacy in our study.

Much attention has been paid to scientific literacy in scientific cognition (Arons, 1983; Branscomb, 1981; Laugksch, 2000; Miller, 1983). To enable individuals to become sufficiently aware of science and science-related public issues, science educators should prepare students to correctly apply scientific knowledge to solve problems and make decisions in their personal, civic, and professional lives. Because this is a part of a big project, other science education professors in Taiwan were responsible for the other five authentic assessments; so only the results of scientific cognition are described in this article. In addition, the members of our research group all majored in chemistry and physics, so we focused on evaluating the concepts of chemistry and physics in the theme of scientific cognition. Comparing our results with test items developed by other researchers (Aikenhead & Ryan, 1992; Laugksch & Spargo, 1996), we propose that scientific cognition could be investigated via different formats of authentic assessments within a limited timeframe. Therefore, the purpose of this article is to display three different formats of authentic assessments to investigate ninth graders' scientific cognition (Figure 1).

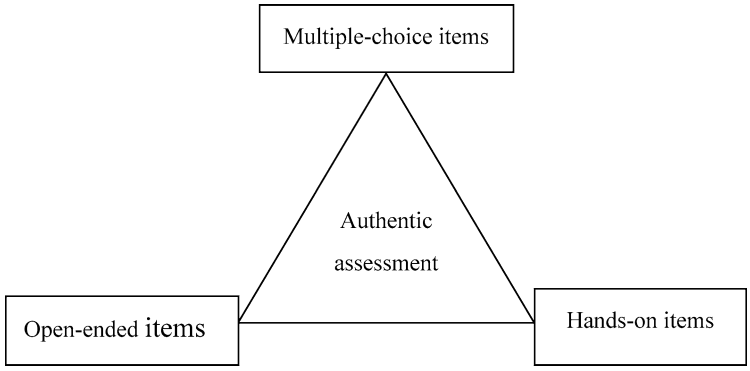


Figure 1. The framework of three authentic assessments.

The purposes of this study were to investigate (1) whether the authentic assessments we developed could be adopted as a national test, and (2) whether scientific cognition could be evaluated by different formats of authentic assessments, which would be better at investigating ninth graders' authentic ability than standardized tests, such as the science portion of the TAAT (STAAT), the well-established high school entrance test currently given in Taiwan. Following on from the research purposes, there were three research questions we wanted to investigate, which were: (1) What were the 9th-graders' performance on the authentic assessments we developed for evaluating scientific cognition? (2) What were the relationships between the three different formats of authentic assessments? (3) How was the students' performance on the hands-on assessment.

METHODOLOGY

In the methodology section, we will describe the research design of our study, the sampling procedure of the participants, and the development and administration of our instruments. Finally the analytic procedure will be presented.

Research Design

The main objective of the study was to develop various authentic assessments as espoused by Cumming and Maxwell (1999), to evaluate ninth graders' scientific literacy. The study described in this article was a two-year project supported by the National Science Council (NSC) and Ministry of Education (MOE) of Taiwan, and the second author of this article was the principle investigator. The procedure of this study was divided into three periods: development period, experiment period, and data analyses

period. During the development period, we developed the core context of the test items and other instruments used in investigating six scientific curriculum themes, and we had the test items validated by experts. The final test items were chosen based on the ratings of discrimination and difficulty from the pre-pilot and pilot study. During the experimental period, we selected the participants by cluster and stratified sampling and then conducted the national test. After that we analyzed the data to find out the effects of our authentic assessments and how students perform on them.

Sampling Procedure

Involved in this study were 1,503 ninth-grade students from 44 schools, their ages averaging 15 years. Subjects were chosen by the method of cluster and stratified sampling. Our basic idea was to randomly select one class from some randomly selected schools. In the first stratum, we chose schools based upon every 2,000 students (the same principle adopted in the TIMSS study). According to the different percentages of student distribution and the different school sizes, 44 schools were randomly selected from north, middle, south, and east of Taiwan. There are three groups of school size in Taiwan, which are large, middle and small sizes. Large size means there are more than 13 classes (including) in 9th grade, middle size represents more than six and fewer than 12 classes (including) in 9th grade, and fewer than five classes (including) in 9th grade is referred to as small size. Because the average number of students per class in Taiwan is 35, we were able to figure out how many classes we needed from the different locations. Then we randomly selected schools and classes from each location. The number of subjects and the size of the schools from which they were chosen are listed in Table II. Due to the limited budget of our study, only 6 of the 44 schools participated in the hands-on test. In total, there were 1,469 students who participated in the multiple-choice and open-ended portions, and 193 students took part in the hands-on portion. The distribution of analysed subjects from different areas and sizes of schools are listed in Table II as well.

Development and Administration of the Instruments

Because national entrance examinations are given within a specified time limit and cover specific main themes, we used these constraints but chose different formats for our assessments. As a research group, we recruited chemistry and physics professors, in-service high school teachers and some graduate students with teaching experience to design paper-and-pencil tests and hands-on assessments to investigate the six themes of scientific literacy. Using the six themes, we developed six kinds of assessments to eval-

TABLE II

Number of student participants, their locations, and the size of their schools

	School size							
	Large size		Medium size		Small size			
	No. of students		No. of students		No. of students		Total No. of students	
Location	Sampling no.	Analysed no.	Sampling no.	Analysed no.	Sampling no.	Analysed no.	Sampling no.	Analysed no.
North	517	494	93	92	21	21	631	607
Middle	264	259	134	134	34	33	432	426
South	271	270	100	99	55	53	426	422
East	–	–	–	–	14	14	14	14
Total	1052	1023	327	325	124	121	1503	1469

Large: if there are more than 13 classes (including) in 9th grade.
Medium: if there are more than 6 and fewer than 12 classes (including) in 9th grade.
Small: if there are fewer than 5 classes (including) in 9th grade.

uate ninth graders' scientific literacy in Taiwan, which included multiple-choice, open-ended and Likert scale items in the paper and pencil tests and hands-on assessments (Table I).

In terms of scientific cognition, we developed multiple-choice items, open-ended items, and a hands-on activity. The 21 multiple-choice items and 4 main open-ended items took the students 60 minutes to complete, and the hands-on activity took the students 30 minutes to complete. The context of the multiple-choice items revolved around five friends who went to a Boy Scout camp in National Yang-Ming Park in Taiwan. While camping, they cooked eggs in the famous hot spring located within the park, prepared dinner by themselves, sat around a campfire at night, and raced cars. The multiple-choice items asked questions about the scientific concepts that may have been present or that the friends might have observed during the activities. The tested concepts will be presented in the following part of results regarding items and themes analysis of authentic assessments.

The open-ended items were, on the other hand, taken from general life experiences, e.g., how alcohol works in cleaning lotion, how to understand the nutrition label on a package of curried chicken, a steam phenomenon that takes place when eating stew, and so on. Regarding the hands-on activity, we asked students to distinguish between four colourless solutions (labeled A, B, C and D in four eyedroppers), by using a universal indicator to detect the pH of these four solutions. During the activity, students were allowed to mix two of these four solutions to observe the reaction between them. In addition to the four eyedroppers, we also provided students with a bottle of universal indicator contained in an eyedropper and a 10 cm $\times$ 10 cm transparency sheet on which to work. This activity would assess students' scientific cognition of pH, the characteristics of these solutions, and the reactions that took place after two of the solutions were mixed. These four solutions were HCl(aq), NaCl(aq), $CaCl_2$(aq) and $CaCO_3$(aq). Examples of the multiple-choice items, open-ended items, and hands-on activity items for understanding students' scientific cognition can be found in Appendix. The reliability (Guttman) of the 21 multiple-choice test items was 0.63 (Lambda 2); of the open-ended questions, 0.80 (Lambda 2); and of the hands-on activity, 0.85 (Lambda 2).

In addition to the test items for investigating students' scientific literacy, we developed supervision and monitoring sheets with instruction for supervisors and monitors. Supervision sheets were to help those teachers responsible for supervising the national test. They had to learn the standard procedure of the test, and the supervision sheets gave them a checklist of points to which they needed to pay attention. Monitoring sheets were

created for monitors who visited 12 schools when the national tests were administered, to make sure the test procedures were correct and to record any special circumstances or occurrences during the test. In addition, we videotaped one class of students while they carried out the hands-on test.

Analytic Procedure

First, we invited some experts from the Departments of Chemistry and Physics at National Taiwan Normal University to validate the concepts of our test items. Then, we conducted the pre-pilot study with approximately 60 ninth-grade students, which led to the elimination of some items by examining their reliability and level of difficulty. Ultimately, we produced 21 multiple-choice test items, 32 open-ended items, and one hands-on activity for the concepts of chemistry and physics. In addition, we conducted the pilot study with around 200 ninth graders to analyze the internal consistency using the Guttman coefficient; this pilot study showed a satisfactory internal consistency of reliability.

To validate the three formats of assessments for scientific cognition, we used Pearson correlation to compare the students' performance with their performance on the science part of TAAT, STAAT. On the one hand, we used MS Access to analyse the consistency of students' performance to see whether students with different achievements in the conventional test, STAAT, would perform consistently in our authentic assessments or not. On the other hand, we used Pearson correlation to see whether the three formats of authentic assessments benefitted the low achievement students.

Regarding the instruments we developed, we made item analyses of our authentic assessments to know what are the difficult concepts for 9th graders based on the average percentages of correct answers concerning specific themes presented in the test items.

RESULTS

For the results section, we will disclose the validity of the three formats of authentic assessments and show the students' performance and relationship between the authentic assessments we developed and the STAAT. Furthermore, we will present what we found in terms of different levels of students' performance on the three authentic assessments. Finally, we will reveal the results from the items and themes analysis of our authentic assessments.

TABLE III

The values of Pearson correlation between the three formats of authentic assessments and STAAT

	Multiple-choice items	Open-ended items	Hands-on items	STAAT
Multiple-choice items	1.00 ($n = 1469$)	–	–	–
Open-ended items	0.660* ($n = 1469$)	1.00 ($n = 1469$)	–	–
Hands-on items	0.437* ($n = 193$)	0.451* ($n = 193$)	1.00 ($n = 193$)	–
Total scores of authentic assessments	0.732* ($n = 193$)	0.753* ($n = 193$)	0.640* ($n = 193$)	0.00 ($n = 193$)
STAAT	0.301* ($n = 1469$)	0.265* ($n = 1469$)	0.205* ($n = 193$)	1.00 ($n = 1469$)

*Correlation is significant at the 0.01 level (2-tailed).
n = total numbers of students participating in each test.

The Validity of Three Formats of Assessments for Scientific Cognition

To validate our three formats of authentic assessments, we compared the performance of 1,469 ninth-grade students on the multiple-choice and open-ended test items, and the performance of 193 ninth graders on the hands-on portion, with the STAAT. The values of the Pearson correlations are shown in Table 3. We found the three authentic tests we developed are all highly correlated with the entrance examination STAAT. The values of the Pearson correlations between students' accomplishments in these tests and the STAAT were all at significant levels, ranging from 0.205 to 0.301 ($p < 0.01$). In terms of the correlations between the three formats of assessments, the hands-on activity test showed high correlations with the multiple-choice (0.437, $p < 0.01$) and open-ended test items (0.451, $p < 0.01$). The value of the correlation between the multiple-choice and open-ended test items was 0.660 ($p < 0.01$) and thus significant. Furthermore, to see the correlations between each format and the total scores of the three formats, we analysed the performance of 193 students who took part in all three formats of authentic assessment (Table III). The values of the Pearson correlation were all at significant levels, from 0.640 to 0.753 ($p < 0.01$).

The Individual Performance of Three Groups of Achievement in the STAAT and Three Authentic Assessments

From these tests, we wanted to know whether individual performance in the STAAT and our three authentic assessments were consistent or not. At first, we divided students into three groups according to their achievements in assessments, whether in the STAAT or our three authentic assessments. The top 27% of students were assigned to the high achievement group, and the lowest 27% of students were assigned to the low achievement group. The rest of the students were in the middle achievement group. Further, we used Access to investigate whether students who belonged to the high achievement group in the STAAT also belonged to the high achievement group in our three authentic assessments. According to the same principle, we analysed the students in the middle and low achievement groups. From our results, we found that 67.4% and 66.7% of the students who were in the high achievement group performed in the top 27% on the multiple-choice and open-ended tests (Figure 2). Similar percentages were discovered in the middle and low achievement groups. As expected, only 40.4% to 46.2% of students in all three-achievement groups of the STAAT performed similarly on the hands-on portion of our assessments. These results show that the standardized test was unable to evaluate students' authentic ability, especially in the hands-on portion.

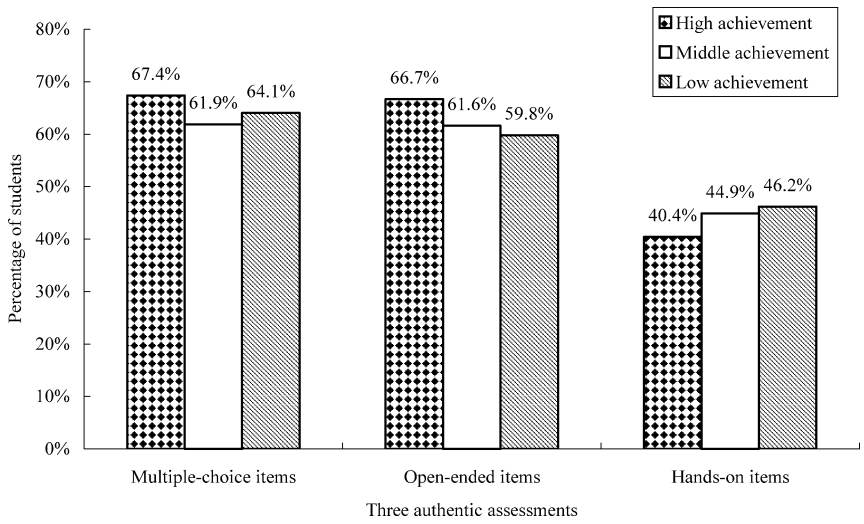


Figure 2. The individual performance of three groups of achievement in the STAAT and the three authentic assessments.

TABLE IV

The values of Pearson correlation between the three groups of achievements in the hands-on test and the total scores of the three formats of authentic assessments

	Hands-on test		
Authentic assessments	High achievement	Middle achievement	Low achievement
Total scores	0.414*	0.273*	0.507*

*Correlation is significant at the 0.01 level (2-tailed).

Authentic Assessment Benefits Low Achievement Students

We were also interested in learning whether a correlation existed between students' hands-on performance and the total scores of our three formats of authentic assessments. We divided the students into three groups according to their total scores on the three formats of our authentic assessments, and compared that with the performance of the three-achievement groups on the hands-on portion. The Pearson correlation was 0.507 ($p < 0.01$) and thus significant for the low achievement group (Table IV).

Items and Themes Analysis of Authentic Assessments

Regarding the items analysis, among the multiple-choice items, we found that students scored lowest (the percentages of correct answers are below 52%) on items related to concepts in electricity: circuit, current, and electrical energy (Items 7–10, 16–21). Meanwhile, we found students scored lower on questions dealing with the concepts most related to phenomena occurring in everyday life (Items 4, 11, and 13). Item 4 was about comparing the conductivity of metal, item 11 was about the three states of water while boiling water (air, vapour and water), and item 13 was about the buoyancy mechanism observed while cooking dumplings.

Concerning the open-ended questions, the students scored lower on those questions regarding the concepts related to everyday phenomena as multiple-choice items, but performed better in questions about the composition of alcohol (item I.1a, with the percentage of correct answers being 51.7%) and air gas (items III.4a, 4b and 4c, with the percentage of correct answers being 54.1%–63.2%). These questions were designed similarly to those seen in students' textbooks. Some students still had misconceptions about alcohol; they thought it was an organic compound, and they were not particularly familiar with alcohol's characteristics. For instance, some students thought alcohol was a strong base that could not dissolve oil.

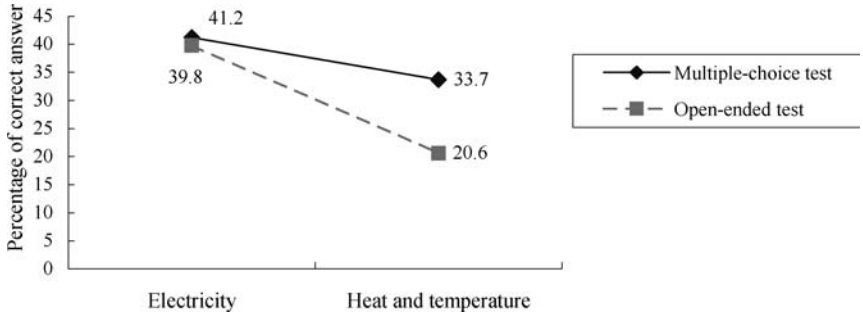


Figure 3. The percentage of correct answer on multiple-choice and open-ended tests regarding the themes of electricity, and heat and temperature.

However, the students performed beyond our expectations on the hands-on activity, with the percentage of correct answer being 73.2%. Students seemed to find it relatively easy to distinguish HCl(aq) from the other three solutions, which were NaCl(aq), $CaCl_2$(aq) and $CaCO_3$(aq).

Furthermore, we compared their performances on the same themes between three formats of authentic assessments. The themes and average percentages of correct answers for these three formats of items are shown in Table V. As mentioned above, electricity, and heat and temperature, were both hard themes for students in our study to grasp. According to the average percentages of correct answers concerning these two themes under multiple-choice and open-ended assessments, the students performed better in the multiple-choice items than in the open-ended ones (Figure 3). However, we found that students were good at electrical calculation, but not good at predicting the lifespan of a battery in a different set of motors (Appendix, item 17). Besides this, we found that the students had misconceptions about the sequence in which bulbs would light in a series connection (Appendix, item 18). On item 17, 63.3% of the students chose answer (a), and on item 18, 50.7% of the students chose answer (a).

Moreover, chemical reactions and the reactions of acid and base with indicators were the only two themes chosen after the pilot study and tested in all three authentic assessments, and we found that the students performed better in the hands-on activity (Figure 4). The students could more easily distinguish an unknown solution as an acid or a base using litmus test paper (item 1 in multiple-choice test) or a universal indicator (hands-on test), but they had difficulty in identifying unknown solutions when presented with the results from two kinds of tests using a litmus test paper and phenolphthalein (items III.2a and III.2b in open-ended test), especially while the solution was neutral.

TABLE V

The average percentage of correct answers on specific themes tested

Themes tested	Percentage of correct answer	Item no.
Multiple-choice items		
Knowing the phenomenon of how indicators react with acid or base.	78.0	1
Knowing the characteristics of chemical reactions.	64.4	2, 3
Heat and temperature.	33.7	4, 5, 11, 14
Living thing's reaction to environmental stimulus and animal behaviours.	61.1	6
Electricity.	52.0	7, 8, 9, 10, 16, 17, 18, 19, 10, 21
The changes of atoms and molecules in a physics reaction from the particle's perspective.	59.6	12
Buoyancy mechanism.	43.2	13
Organic compounds.	10.5	15
Open-ended items		
The concept of carbohydrate.	51.7	I.1a
Solution is a composition of a solute on a solvent.	6.6	I.1b
The concept of evaporation.	18.5	I.1c, I.1d, I.2d
Calculation of calories.	39.0	I.2a
The concepts of food processing and vacuum packaging of food.	16.1	I.2b, I.2c
Heat and temperature.	20.6	II.1
Knowing the influence of pressure on gas.	16.3	II.2
Knowing the reflection of light.	15.0	II.3
Calculation of electricity.	39.8	II.4
Knowing how to change the representation of data.	82.5	III.1a
Knowing how to find a rule of a data set to calculate the volume of gas generated from electrolysis of water.	0.1	III.1b
Knowing how to change the representation of a molecule.	7.6	III.1c, III.1d
Knowing the phenomenon of how indicators react with acid or base.	19.6	III.2a, III.2b
Knowing the characteristics of chemical reactions.	22.2	III.3

TABLE V

(Continued)

Themes tested	Percentage of correct answer	Item no.
The concept of the composition of air.	57.4	III.4a, III.4b, III.4c
Judging the speed of a car by the distance and time elapsed.	41.3	IV.1, IV.2a, IV.2b, IV.2c
Hands-on activity items		
Knowing the phenomenon of how indicators react with acid or base.	82.2	I.1, I.2, I.3, I.4, I.5, I.6, I.7, I.8
Knowing the characteristics of chemical reactions.	79.3	II.A–D, II.B–D, II.C–D, II.A–C, II.B–C, II.A–B
Could conclude the name of the solution from the data.	58.1	III.1a, III.1b, III.2a, III.2b, III.3a, III.3b, III.4a, III.4b

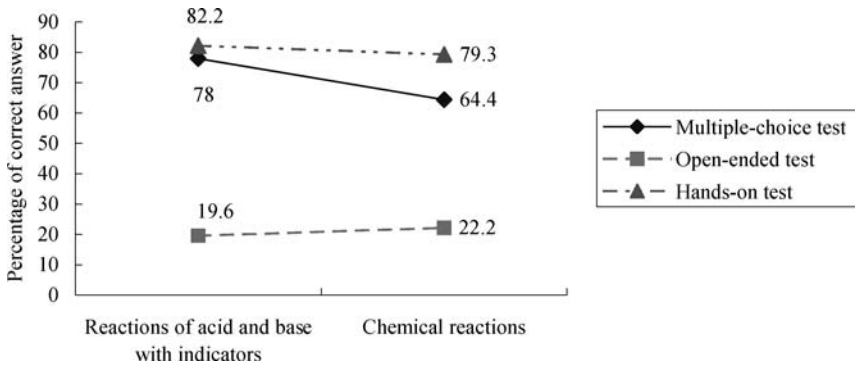


Figure 4. The percentage of correct answer on the three formats of authentic assessments regarding the themes of reactions of acid and base with indicators and chemical reactions.

DISCUSSION

The two-year project attempted to develop authentic assessments, which could be adopted as national tests to investigate students' scientific literacy. The assessments described in this article may serve as evidence that authentic assessments could be developed in various formats for the national test, taking into account the time limitation, to investigate students'

scientific literacy. Concerning the development and administration of the instruments, even though it took us a long time to develop test items in the development stage, the students' performance and the acceptance from teachers and students brought positive feedback. In particular, the brief interviews with students and the on-site monitoring by supervisors showed that the students found it interesting to take the open-ended assessment, and most of students performed the hands-on activity with smiles. From the monitoring and supervision sheets, although we found there were a few teachers who did not follow the standard procedures (about 3% of the teachers were not familiar with standard procedures), it did not influence the implementation of the assessments. In general, the supervisors all agreed on the good quality of the execution of the national test.

The analytic results of this study regarding scientific cognition indicated that the three formats of assessments we developed were validated according to the high correlations with each other and the STAAT. According to the individual performance of the three groups of achievement, the authentic assessments were better than the standardized test, STAAT, in terms of its ease for probing students' authentic ability, since only around 60% of the three-achievement group students from the STAAT performed consistently in both multiple-choice and open-ended tests. In addition, only around 40% of the three-achievement group students from the STAAT performed consistently in the hands-on test (Figure 2). In fact the three formats of authentic assessment tested students' authentic ability differently in the different achievement groups. In spite of the fact that the three authentic assessments could each probe students' authentic ability, our hands-on test definitely benefitted the low achievement group students, based on the evidence provided by Pearson correlation (Table IV). That is to say in the low achievement group, based on total scores, the hands-on test contributed more scores than the other two formats of authentic assessments.

From the results of individual items and analysis of themes, the students scored lower in chemistry when dealing with real-life phenomena, and this was a common issue (Walford, 1983). In the meantime, the ninth-grade students scored lowest on the items related to concepts in electricity, a finding which corresponds to the many studies that reveal that electricity concepts are difficult for children to learn, especially the concepts of circuit, current, and electrical energy (Shipstone, 1988; Garnett & Treagust, 1992; Furio & Guisasola, 1998). We discovered two common misconceptions regarding electricity in our study, which have also been revealed by many researchers (Magnusson, Boyle & Templin, 1997; Osborne, 1983; Russell, 1980; Shipstone, 1984): Students thought that the bulb in a series

nearest the positive electrode would light first (Appendix, item 18), and that two bulbs together use more electricity than a single bulb (Appendix, item 17).

Furthermore, we explored the content in terms of chemical reactions and electricity in the textbooks used in junior high school. The presentation of hands-on activities and embedded scientific concepts simultaneously meant that the students could know the results of the activities from the following content without thinking by themselves. In fact, due to the time limitation of the lessons, teachers might skip the hands-on activities and conduct demonstrations instead. Besides the arrangement of hands-on activities in the textbooks and during lessons being a potential problem, the content of the activities are too scientific to transfer. For example, regarding the theme of heat and temperature, the equipment needed are thermometers, cone bottles and alcohol burners to let students observe the change of temperature and conductivity of heat. This seems to lack the connections which the phenomenon might have in real-life, and lead to the difficulty in transferring scientific concepts to life beyond school. Because of the time limitations of the national test, we could not evaluate all the themes taught in junior high school by the instruments. However, it is a work we could endeavour to develop in the future.

Conclusions and Implications

In this new century, many countries not only focus on research in science and technology, but also dedicate time and resources to the improvement of science education. Education reform has become a continuous movement in many countries (Bell et al., 2001; Education Act 1996, 1996; McComas et al., 1998), and Taiwan is following the same trend. In the current science education and reform movement, authentic science has been considered an important perspective, and authenticity has been identified as an emergent property of science learning (Rahm, Miller, Hartley & Moore, 2003). It is exactly the spirit of scientific literacy. According to Miller's opinion, scientific literacy in today's society should help students understand the norms and methods of science, key scientific terms and concepts, and the impact of science and technology on society beyond school (Miller, 1983). The main contribution of our current study is the development and validation of the various authentic assessments to evaluate ninth graders' scientific literacy, and thus serve as the evidence of implementation as a national test.

The notion of authenticity has led many researchers to emphasize the importance of curriculum design for authentic science (Cunningham &

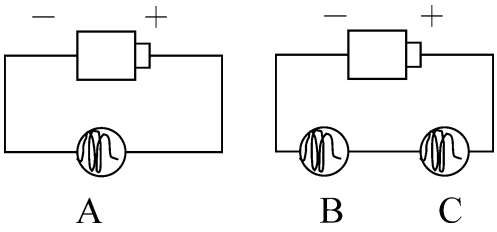
Helms, 1998; Driver, Asoko, Leach, Mortimer & Scott, 1994) and the development of authentic assessment (Cumming & Maxwell, 1999; Dori, 2003; Wiggins, 1989, 1993; Wolf et al., 1991). Such assessment is sensitive enough to evaluate students' abilities or learning outcomes, and could reflect students' deep and broad understanding (Dori, 2003). Currently, standardized and authentic testing have started to merge for evaluating students' higher-order thinking skills, which teachers have tended to neglect when using test-oriented teaching (Dori, 2003). Therefore, we set out to develop authentic assessments to explore students' scientific literacy in our two-year project. In this article, we provide evidence that authentic assessment could be developed in different formats to investigate students' scientific literacy as part of the national test and to display an effective way of evaluating students' authentic abilities of science.

What should we do as science educators as the next step? We need to know the student's misconceptions and difficulties regarding our test items via interviewing students. There has been some research conducted, thus far, on improving students' learning in electricity (Lee & Law, 2001; Shipstone, 1988; Steinberg & Wainwright, 1993) and chemistry (Binns, 1978; Gabel & Sherwood, 1984; Jegl, 1978; Kilker, 1985; Sevenair, 1989). We need to add to that research base by designing models or using computer-based programs to bridge the gap between school learning and real life, in order to improve the learning of difficult concepts. We also must provide the results to in-service teachers (1) so that they will understand the difficulties that exist in student learning and (2) to build up a long-term cadre of teachers who can share the effective teaching models with other in-service teachers. Furthermore, we need to be vigilant about making sure that in-service teachers pay attention to whether students are reflecting on real-life phenomena, rather than having students rely solely on the scientific knowledge they learn in the classroom (Yerrick, 2000). In other words, learning scientific concepts should be embedded in context for making science learning more meaningful (Rahm et al., 2003).

ACKNOWLEDGEMENT

The authors wish to thank the National Science Council and Ministry of Education of Taiwan, the Republic of China, which has financed the two-year project (grant no. NSC89-2515-S-003-012-X3 and NSC90-2511-S-003-101-X3). The authors also acknowledge the assistance of Chain-Huey Yang, Pey-Li Leong, Yi-Ju Lin, and Jin-Wen Lin when implemented the project.

APPENDIX. THE EXAMPLES OF ASSESSMENTS REGARDING SCIENTIFIC COGNITION.

Format of the items		Contents	Aspects tested
Multiple-choice items		Tomorrow morning, there will be a car race, so John wants to set up the lights in his car so they are best suited for the competition. Now he has the same batteries, light bulbs, and wires for pre-test:	Electricity
	Item 17	( ) During the competition, which set of batteries shown below could be used longer? (a) the set with bulb A; (b) the set with bulbs B and C; (c) the batteries in both sets could be used for the same amount of time. Set I Set II  	
	Item 18	( ) Which bulb do you think will light first after John connects the wires in set II? (a) C; (b) B; (c) B and C will light at the same time.	
Open-ended items	Item II.1	While enjoying some stew at home, you notice that a mist covers father's glasses. Can you explain the reason for this phenomenon? Answer: _____ _____	Knowing how matter is influenced by temperature

Hands-on activity items	Mary put four kinds of colorless solutions (HCl(aq), NaCl(aq), $CaCl_2$(aq) and $CaCO_3$(aq)) into four bottles, but she forgot to write down the names of the solutions on the bottles (they are labelled only A, B, C and D). Can you help Mary to distinguish these four kinds of solutions in the bottles by using a universal indicator? The table below shows the colours presented for different pH values.	Knowing the characteristics of these solutions, the phenomenon of acid or base react on a universal indicator, and the reactions happened between both of the solutions.

pH	Colour
1	Red
3	Orange
5	Yellow
7	Green
9	Blue
11	Purple
13	Black

(I) Write down the colours observed after using the universal indicator and then distinguish the acid or base of these four solutions in the table below:

Labels	A	B	C	D
Colours				
Acid or base				

(II) Mix the four solutions in groups of two, and, then write down the phenomenon you observed.

	A	B	C
D			
C			
B			

(III) From the activities that you've carried out before, write down what solutions are in of the four bottles, and give your reasons.

The solution in A bottle is ____, because _____.

The solution in B bottle is ____, because _____.

The solution in C bottle is ____, because _____.

The solution in D bottle is ____, because _____.

REFERENCES

American Association for Advancement of Science (AAAS) (1993). *Benchmarks for Science Literacy*. Washington, DC: AAAS.

Aikenhead, G.S. & Ryan, A.G. (1992). The development of a new instrument: "Views on science-technology-society" (VOSTS). *Science Education, 76*(5), 477–491.

Archbald, D.A. & Newmann, F.M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. Reston, VA: National Association of School Principals.

Arons, A.B. (1983). Student patterns of thinking and reasoning: Part one of three parts. *Physics Teacher, 21*(9), 576–581.

Bell, R., Abd-El-Khalick, F., Lederman, N.G., McComas, W.F. & Matthews, M.R. (2001). The nature of science and science education: A bibliography. *Science Education, 10*(1–2), 187–204.

Binns, M. (1978). Chemistry for life: A mode III course. *Education in Chemistry, 15*(5), 143–145.

Branscomb, A.W. (1981). Knowing how to know. *Science, Technology, and Human Values, 6*(36), 5–9.

Champagne, A.B. & Newell, S.T. (1992). Directions for research and development: Alternative methods of assessing scientific literacy. *Journal of Research in Science Teaching, 29*(8), 841–860.

Chiu, M.H. (2002). *The development of assessments of science learning for junior high school curriculum.* No. NSC 90-2511-S-003-101-X3. Taipei, Taiwan: MOE.

Cumming, J.J. & Maxwell, G.S. (1999). Contextualising authentic assessment. *Assessment in Education, 6*(2), 177–194.

Cunningham, C.M. & Helms, J.V. (1998). Sociology of science as a means to a more authentic, inclusive science education. *Journal of Research in Science Teaching, 35*, 483–499.

Dori, Y.J. (2003). From nationwide standardized testing to school-based alternative embedded assessment in Israel: Students' performance in the matriculation 2000 project. *Journal of Research in Science Teaching, 40*(1), 34–52.

Driver, R., Asoko, H., Leach, J., Mortimer, E. & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23*, 5–12.

Education Act 1996 (1996). London: Her Majesty's Stationery Office.

Furio, C. & Guisasola, J. (1998). Difficulties in learning the concept of electric field. *Science Education, 82*(4), 511–526.

Gabel, D. & Sherwood, R.D. (1984). Analyzing difficulties with mole-concept tasks by using familiar analog tasks. *Journal of Research in Science Teaching, 21*(8), 843–851.

Garnett, P.J. & Treagust, D.F. (1992). Conceptual difficulties experienced by senior high school students of electrochemistry: Electric circuits and oxidation–reduction equations. *Journal of Research in Science Teaching, 29*(2), 121–142.

Jegl, W. (1978). The chemistry of life: A second semester course on color videotapes for students in life sciences. *Journal of Chemical Education, 55*(4), 225–259.

Jenkins, E.W. (1992). School science education: Towards a reconstruction. *Journal of Curriculum Studies, 24*(3), 229–246.

Kilker, R.J. (1985). A chemistry course for high ability 8th, 9th, and 10th graders. *Journal of Chemical Education, 62*(5), 423–424.

Laugksch, R.C. (2000). Scientific literacy: A conceptual overview. *Science Education, 84*, 71–94.

Laugksch, R.C. & Spargo, P.E. (1996). Development of a pool of scientific literacy test items based on selected AAAS literacy goals. *Science Education, 80*(2), 121–143.

Lee, Y. & Law, N. (2001). Explorations in promoting conceptual change in electrical concepts via ontological category shift. *International Journal of Science Education, 23*(2), 111–149.

Magnusson, S.J., Boyle, R.A. & Templin, M. (1997). Dynamic science assessment: A new approach for investigating conceptual change. *The Journal of the Learning Sciences, 6*(1), 91–142.

McComas, W.F., Almazroa, H. & Clough, M.P. (1998). The nature of science in science education: An introduction. *Science Education, 7*(6), 511–532.

Miller, J.D. (1983). Scientific literacy: A conceptual and empirical review. *Daedalus, 112*(2), 29–48.

Ministry of Education (MOE) (1998). *1–9 grades curriculum guidelines*. Taipei: MOE.

Newmann, F.M. & Archbald, D.A. (1992). The nature of authentic academic achievement. In H. Berlak, F.M. Newmann, E. Adams, D.A. Archbald, T. Burgess, J. Raven & T.A. Romberg (Eds.), *Toward a new science of educational testing and assessment*. Albany, NY: State University of New York Press.

Osborne, R. (1983). Towards modifying children's ideas about electric current. *Research in Science and Technological Education, 1*(1), 73–82.

Rahm, J., Miller, H.C., Hartley, L. & Moore, J.C. (2003). The value of an emergent notion of authenticity: Examples from two student/teacher-scientist partnership programs. *Journal of Research in Science Teaching, 40*(8), 737–756.

Russell, T.J. (1980). Children's understanding of simple electrical circuits. In A.P.C. Sia (Ed.), *Science and mathematics concept learning of South East children: Second report on Phase II* (pp. 67–91). Malaysia: SEAMEO-RECSAM.

Sevenair, J.P. (1989). A nontraditional organic chemistry course. *Journal of College Science Teaching, 18*(4), 236–239.

Shipstone, D.M. (1984). A study of children's understanding of electricity in simple DC circuits. *European Journal of Science Education, 6*(2), 185–198.
Shipstone, D.M. (1988). A study of students' understanding of electricity in five European countries. *International Journal of Science Education, 10*(3), 303–316.
Shortland, M. (1988). Advocating science: Literacy and public understanding. *Impact of Science on Society, 38*(4), 305–316.
Steinberg, M.S. & Wainwright, C.L. (1993). Using models to teach electricity: The CASTLE Project. *Physics Teacher, 31*(6), 353–357.
Thomas, G. & Durant, J. (1987). *Why should we promote the public understanding of science?* Oxford, UK: University of Oxford.
Walford, E.T. (1983). High school chemistry: Preparation for college or preparation for life? *Journal of Chemical Education, 60*(12), 1053–1055.
Wiggins, G.P. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703–713.
Wiggins, G.P. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*, 201–214.
Wolf, D., Bixby, J., Glenn, J.I. & Gardner, H. (1991). To use their mind well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31–73.
Yerrick, R.K. (2000). Lower track science students' argumentation and open inquiry instruction. *Journal of Research in Science Teaching, 37*(8), 807–838.

Graduate Institute of Science Education,
National Taiwan Normal University,
88, Sec. 4, Ting-Chou Rd., Taipei,
Taiwan 116
E-mail: mhchiu@cc.ntnu.edu.tw