International Journal of Innovative
Computing, Information and Control
Volume **11**, Number **6**, December **2015**

ICIC International ©2015 ISSN 1349-4198
pp. 2193–2213

ANALYZING A COMPUTERIZED DIAGNOSTIC TEST WITH MULTIPLE CHOICE ITEMS AND CONSTRUCTED RESPONSE ITEMS BASED ON BAYESIAN NETWORK

Shih-Hsin Li[1], Nai-Fen Gu[2], Kenneth Lian Lee[3], Bor-Chen Kuo[1] and Chih-Wei Yang[1]

[1]Graduate Institute of Educational Measurement and Statistics
National Taichung University of Education
No. 140, Minsheng Rd., West Dist., Taichung 40306, Taiwan
shihshunlee@gmail.com; kbc@mail.ntcu.edu.tw; chihwei_yang@hotmail.com

[2]Chung-Lun Junior High School
Taichung 402, Taiwan
lubawe@gmail.com

[3]Dadu Elementary School
Taichung 432, Taiwan
jely08330325@yahoo.com.tw

Received April 2015; revised September 2015

Abstract. *Conceptual knowledge and procedural knowledge both influence the construction of students' misconceptions and sub-skills of science. However, there are very few researches which aimed at automatic diagnosis of students' misconceptions and sub-skills through these two types of knowledge together. The purposes of this study were to develop a computerized diagnostic test system to diagnose examinees' misconceptions and sub-skills and score their answering responses to multiple choice items and constructed response items together in the unit of "Air and Combustion" automatically and accurately. In this study a computerized diagnostic test composed of 23 multiple choice items and 3 constructed response items in the context of science curriculum in junior high school was designed and administered to 577 fifth graders in Taiwan. For automatic diagnoses and scoring, a diagnostic model based on Bayesian network was designed and embedded in the aforementioned test system for further immediate analysis of students' complicated answering responses. The results showed that overall immediate diagnostic accuracies of students' misconceptions and sub-skills were both above 90%, which implied that the proposed diagnostic mechanism could assess students' misconceptions and sub-skills automatically and accurately. In addition, some students' misconceptions which could not be assessed only by 23 multiple-choice items were diagnosed by the 3 constructed response items in this test design.*

Keywords: Misconceptions, Sub-skills, Constructed response items, Bayesian network

1. **Introduction.** An important point of students' science learning is the early detection and complete understanding of students' critical bottlenecks to science learning in classes. Therefore, it is necessary to collect and analyze subtle, crucial and complicated data from students' cognitive processes and responses. Thence, educators and researchers make efforts to assess those data and try hard to analyze and transfer them into useful information for educational purposes, for the reason that there is a determinant need to develop assessments for diagnostic purposes. In order to achieve this goal, it is also essential to develop effective testing methods which could correctly estimate examinee's science proficiency and then provide detailed learning information to teachers based on it.

However, the currently existing testing methods such as paper-based testing are inadequate in modern testing need because this kind of method not only need time-consuming testing task but also is often troubled with inadequate item formats. The requirements for reducing the guessing effects and assessing more complete answering responses arouse the motivations to design and evaluate an effective computerized test system with various formats of test items for diagnostic purposes.

Fortunately, the fast development of technology and human interaction in today's world provides the opportunities to change the item formats in the educational measurement [1]. Through a computerized test, students are not limited to just selecting one of several available response alternatives. They can click on buttons, highlight texts, drag or move objects (graphics) around the screen, or re-order a series of statements or pictures [2-4]. Moreover, the computer's ability to interact with students provides more possible forms of interaction between test items and examinees. Items are not only restricted to merely accepting students' answering responses for scoring but also possible for further automatic diagnosis [5-9].

1.1. **The formats of items.** As to the item format, the multiple choice (MC) items involve four or five answer alternatives. Various solutions from which they have to choose the correct answer to a particular question are shown to students. The forms of constructed response (CR) items range from short sentences, phrases to an answer, lengthy essay writings completion of given tasks or generate answers [10-12]. For assessing students' different kinds of abilities/skills, assessment instruments could probably be composed of a combination of these two types of items [13,14].

As to the assessment of students' science learning, the major value of CR item is that it requires students to create (develop) their own responses rather than choose a prepackaged response from the answer shelf. Obviously, creating a response represents a much more complicated and deeper-thinking cognitive processes. Although CR items are tougher for test-takers and much more time-consuming for test-designers and teachers, educators still lay more and more emphases on the benefits of them. Furthermore, assessments with CR items are more convenient for assessing students' higher levels of thinking, yet they are much more difficult and time-consuming to score.

All in all, we can conclude that the design and implementation of tests with CR items are arduous for teachers and test administrators. As a matter of fact, it is indeed not an easy task to put the CR items into practice in an achievement test, let alone in a diagnostic test.

1.2. **The current applications of CR items.** CR items can measure the students' thinking, problem solving processes, and organizational integration and expression skills. There are large scale tests that included CR items, such as National Assessment of Educational Progress (NAEP) [15,16], the trends in International Mathematics and Science Study (TIMSS) [17,18] and the Programme for International Student Assessment (PISA) [19-21]. CR items can detect complicated answering responses provided by examinees, which were manually given different scores by teachers or test administrator [22-27]. Nevertheless, to design and implement a test with CR items takes up a lot of manpower and time, and it is not easy to have immediate feedback [28].

Although CR items can more easily assess higher levels of students' thinking, they are still very difficult to score even by a computerized system. Taking scantrons (optical grade scanners) for example, computers have difficulties for analyzing and scoring these complicated types of answering responses from CR items. Obviously, the computer-based scoring mechanism for CR items is still difficult because the analytic procedure must give serious consideration to all possible kinds of students' answering responses, much less the

computer-based scoring mechanism for the combination of MC items and CR items. And this perplexity is one of the major obstacles we try to overcome.

1.3. **Artificial intelligence for computerized diagnostic test.** As the development of computer technologies, a computerized-based test (CBT) could be composed of both MC items and CR items [23,29-31]. For instance, OECD [32] advocated the use of CBT in the Programme for International Student Assessment (PISA) in science, and indicated that CBT is particularly useful in the assessment of science for simulating scientific phenomena [33,34]. Moreover, teachers and researchers may design the interfaces to test-takers by augmented reality through multimedia [35].

However, it is still not easy for designing an effective and valid computerized diagnostic test mechanism for scoring or grading students' complicated and various responses of CR items. For example, scantrons (optical grade scanners) could not score them. In other words, educators or test administrators need a valid and reliable computer-based scoring rubric and mechanism to differentiate and analyze all kinds of analogous and acceptable answering responses to CR items.

In order to break through the aforementioned barrier, some scholars successfully apply Bayesian network (BN) on educational assessments and design evidence-centered BN structure by categorizing the answering responses of examinees into test data and training data [36-40]. BN is a probabilistic graphical model capable of modeling the certain domain knowledge comprising uncertainties [41]. BN encodes and illustrates qualitative and quantitative parts of domain knowledge by means of a directed acyclic graph $G = (V, E)$, to each node $i \in V$ corresponds one random variable X_i with a finite set x_i of mutually exclusive states for qualitative part, and a conditional probability table (CPT) $P\left(X_i|(X_j)_{j\in pa(i)}\right)$ where $pa(i)$ denotes the set of parents of node i in graph G for quantitative part. In a word, the set P defines the joint probability distribution as:

$$P(X) = \prod_{i=1}^{n} P\left(X_i|(X_j)_{j\in pa(i)}\right) \tag{1}$$

Any desired probabilistic information with a given Bayesian network is obtained by means of Equation (1). As we know, BN has a solid theoretical foundation, equipped with flexible inference capability, and allows human knowledge to be directly encoded into the network. BN has been employed to inference and diagnose various aspects of students' learning [38,40,42,43]. Furthermore, the educational assessment has more flexibility and can obtain broader messages from the diagnostic tests based on BN statistical analysis [42,44-49].

1.4. **Misconceptions in science learning.** A misconception is defined as a perception of phenomena occurring in the real world which is not consistent with the scientific explanation to the phenomena [50]. Methods used to collect students' misconceptions recently are thinking aloud, contextual tests, interviews, concepts analysis, and computerized diagnostic tests [51-55]. However, most of aforementioned methods require lots of manpower and time during the scoring and diagnostic processes. In order to improve these limitations, we lay emphasis on the importance of developing a computerized diagnostic test embedded with BN algorithm which could diagnose students' misconceptions and skills automatically and precisely.

As we know, some studies reveal that students have developed various misconceptions in the unit of combustion [56-58]. Thence, there is the practical requirement for us to choose the unit of combustion to be the research content for educational considerations.

As to the diagnostic purpose, some researches focused on the computerized scoring mechanism of tests with CR items [59,60], yet there are still rare academic references showing that a computerized system with MC items and CR items could diagnose students' misconceptions and skills simultaneously and precisely.

Based on the educational demands to assess and diagnose students' misconceptions and skills from students' dichotomous answering responses from MC items and complicated solving procedures from CR items, we try to use BN known for its powerful knowledge representation to diagnose students' latent traits (misconceptions, skills).

1.5. **Science skills and science knowledge in learning.** Student-Centered teaching and learning approaches are widely emphasized and encouraged in the learning activities and related styles of assessments [61,62]. From the Student-Centered points of views, the development of students' skills for solving the problems of science guide teachers to analyze and separate these skills into smaller elements considering the related scientific content [63]. These so-called smaller elements of skills are defined as sub-skills in this paper.

The sub-skills of students are the basic skills required in solving the problems of learning domain [64,65]. As to students' scientific learning, misconceptions and sub-skills are not isolated from each other but should be considered together [65-68]. By assessing and understanding students' misconceptions and sub-skills, teachers can provide proper learning activities to facilitate explicit, adaptive, and well-planned learning opportunities that foster students construct and develop science skills and knowledge.

As to science knowledge, recent literature reveals the importance of conceptual knowledge and procedural knowledge in the individual development of science knowledge [69-73]. In combination, these two kinds of science knowledge define what students are expected to learn in science class.

From above, we try to propose a computerized diagnostic test system based on BN for diagnosing students' misconceptions and sub-skills from their dichotomous answering responses (right and wrong) from MC items and various patterns of complicated problem solving procedures from CR items by assessing their problem solving strategies. In this paper, the purposes of this research are as the following:

(i) Analyze misconceptions and sub-skills related to the unit of "combustion" of science and life technology learning domain;
(ii) Construct the diagnostic and scoring model with MC items and CR items based on BN for assessing students' misconceptions and sub-skills from purpose (i);
(iii) Implement the computerized test system based on the diagnostic and scoring model from purpose (ii), and evaluate the diagnosis accuracies of CR items, misconceptions and sub-skills.

All in all, there are three important meanings from the academic points of views. First, by the design and implementation of the proposed computerized diagnostic test system with different formats of items, teachers may have the chances to assess students' misconceptions and sub-skills which cannot be easily assessed by MC items. The information can guide teachers to offer proper remedial curriculums for students. Secondly, we try to overcome the greatest problem in test with CR items which are the time and expense involved in scoring. Through BN mechanism embedded in the diagnostic test system in this paper, we can improve and even prevent that scoring process requires substantial amounts of time from highly trained scorers. Thirdly, we try to elaborate and evaluate the consistency and accuracy of the diagnostic and scoring results.

As to the future application teacher educators, curriculum specialists, cognitive psychologists, and researchers may have standard procedures which can effectively and precisely

assess and focus on students' science conceptual knowledge and procedural knowledge (cognitive processes) together based on the following methodology and later results.

2. **Methodology.** After collecting and analyzing students' misconceptions, we established the computerized diagnostic test system composed of 27 MC items and 3CR items based on BN for assessing students' misconceptions and sub-skills in the unit of "combustion" of science and life technology learning domain. Afterwards, the diagnosis accuracy of the proposed computerized test system is evaluated and discussed. The research procedures are as Figure 1.

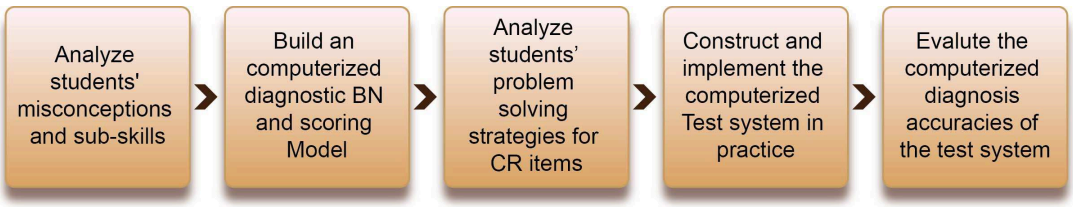

FIGURE 1. Research procedure

2.1. **Learning materials.** The learning materials are designed for the 5th grade students in the first semester according to the unit of "Air and Combustion" of Kang Hsuan[1] Educational Publishing in Taiwan.

2.2. **Participants.** The computerized diagnostic test is targeted for the 5th grade students who have finished learning materials in the unit of "Air and Combustion" of science and life technology learning domain. The total valid samples are 577 students who come from 24 classes of 3 schools in Taichung City of Taiwan.

2.3. **System interface and database format of CR item.** The system interface is divided into 4 areas (see Figure 2); the timer shows the reaction time. Examinee operates the mouse by dragging experimental apparatus into the workspace, pressing "redo" to return to the previous stage. The description and initial state of CR item 1 is shown in Figure 3.

One example of the detailed problem solving processes of CR item 1 is listed as follows: (Step ①) drag a conical flask into the workspace→(Step ②) drag a funnel into the workspace and place it on top of the conical flask→(Step ③) pour hydrogen peroxide through the funnel into the conical flask→(Step ④) drag a funnel into workspace but not place it on top of the conical flask→(Step ⑤) pour diced carrots into the conical flask directly→(Step ⑥) drag a glass plate into the workspace and place it on the conical flask→(Step ⑦) drag an incense into the workspace but not place it on top of the conical flask→(Step ⑧) drag an incense into the workspace and place it on top of the conical flask→(Step ⑨) press the OK button (see Figure 4 to Figure 12).

As to the aforementioned detailed problem solving processes, the corresponding database format is encoded as the following one: BO1_LO1_OO1_LO4_HU2_GL1_GL4_LE1_right. Every possible solving process of each step in CR item 1 is given a code, as shown in Table 1.

[1]Kang Hsuan Educational Publishing Group is one of the professional elementary school K1-K6 textbook publishers in Taiwan.

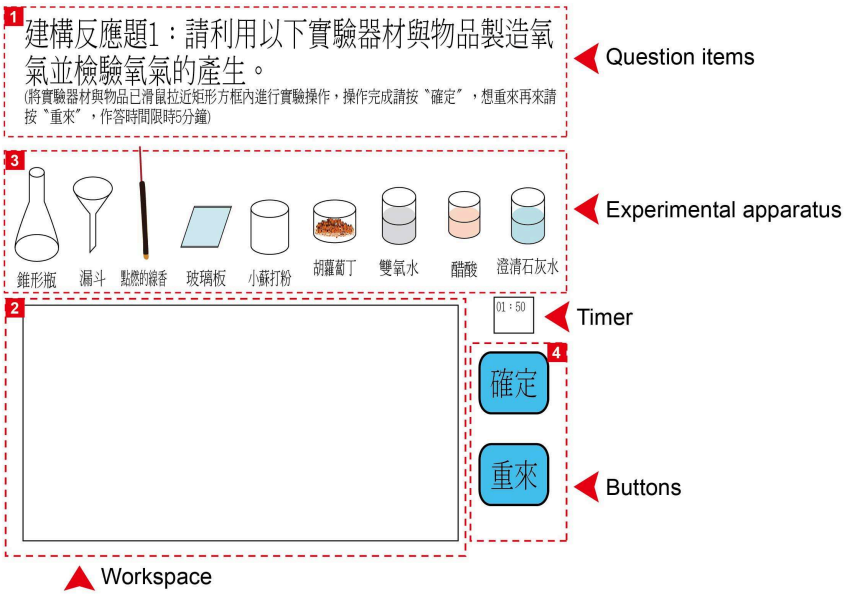

FIGURE 2. Example of system interface

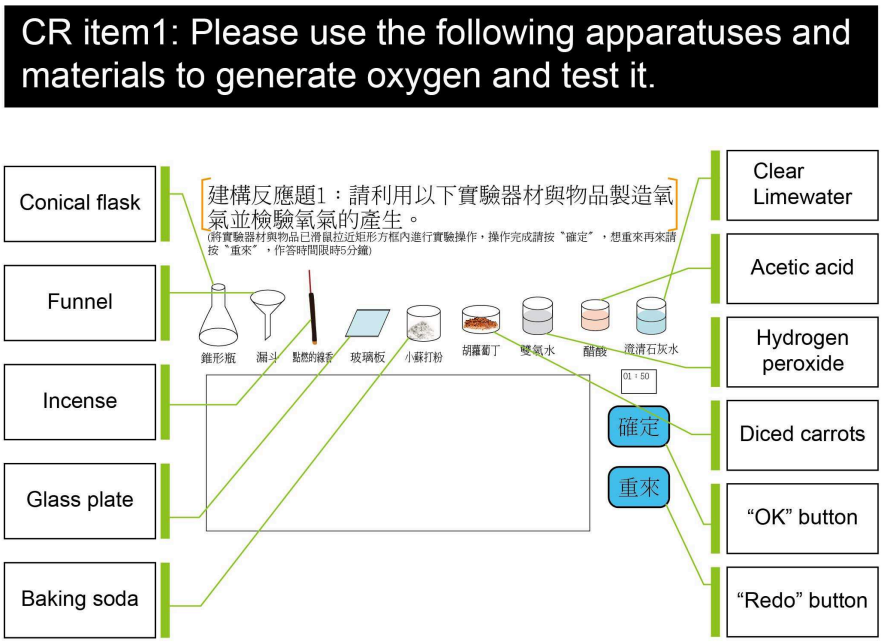

FIGURE 3. The description and initial state of CR item 1

2.4. **Problem solving strategies and scoring rubrics.** This test is composed of 23 MC items and 3 CR items. Due to the paper length limitation, we choose and illustrate one representative example of problem solving strategies and scoring rubrics of MC item 1 and CR item 1 as follows.

2.4.1. *Example of MC items.* Table 2 shows an example of the checklist of the MC item 1 of which the correct alternative was the 3rd one. The student shows that he/she does not construct (develop) related misconceptions (b5, b16), when he/she chooses the 3rd alternative. On the other hand, when he/she chooses another alternative (alternative 1, alternative 2 or alternative 4), he/she may construct (develop) related misconception.

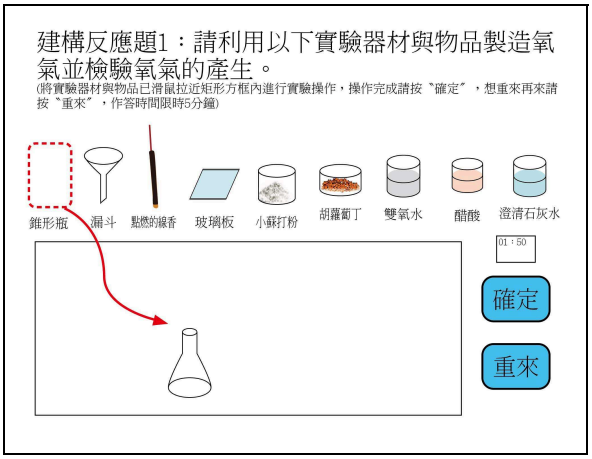

FIGURE 4. Drag a conical flask into the workspace (Step ①)

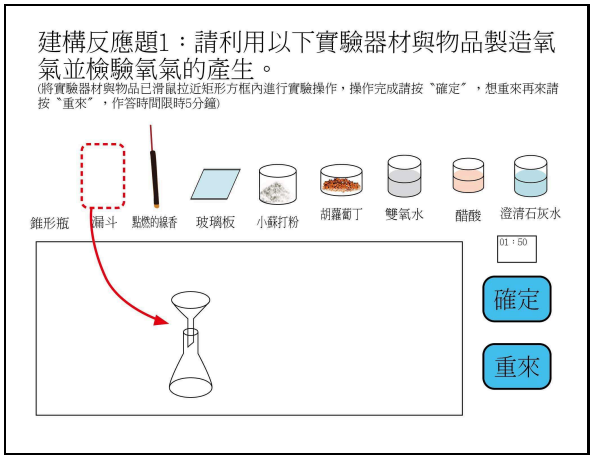

FIGURE 5. Drag a funnel into the workspace and place it on top of the conical flask (Step ②)

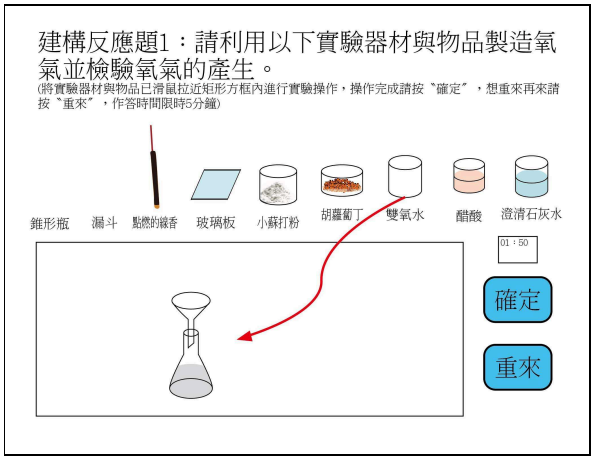

FIGURE 6. Pour hydrogen peroxide through the funnel into the conical flask (Step ③)

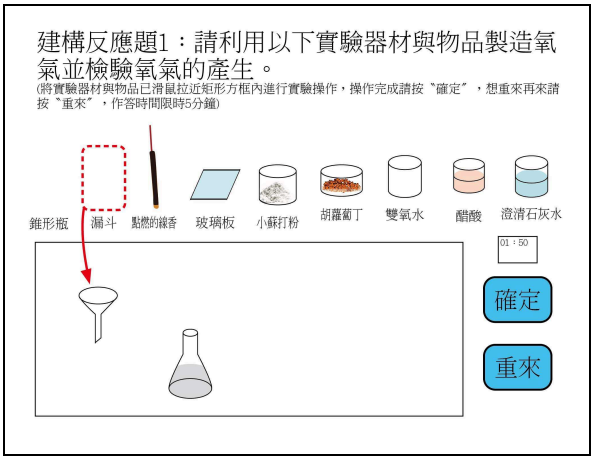

FIGURE 7. Drag a funnel into the workspace but not place it on top of the conical flask (Step ④)

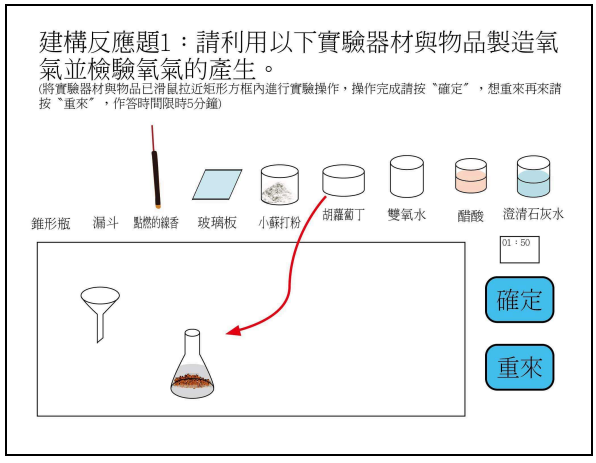

FIGURE 8. Pour diced carrots into the conical flask directly (Step ⑤)

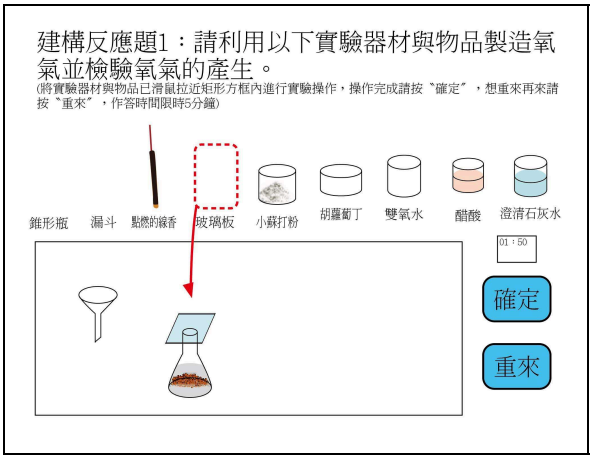

FIGURE 9. Drag a glass plate into the workspace and place it on the conical flask (Step ⑥)

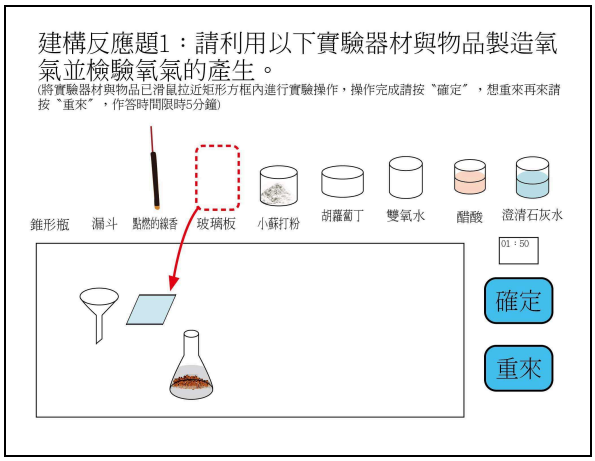

FIGURE 10. Drag a glass plate into the workspace but not place it on top of the conical flask (Step ⑦)

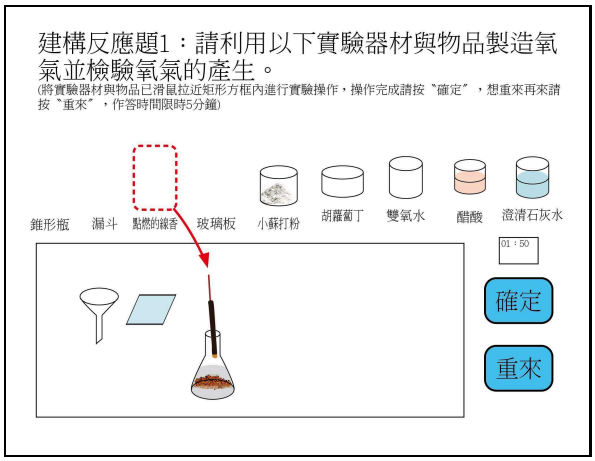

FIGURE 11. Drag an incense into the workspace and place it on top of the conical flask (Step ⑧)

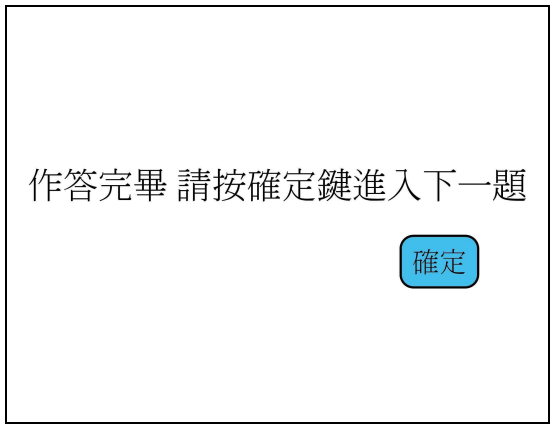

FIGURE 12. Press the "OK" button (Step ⑨)

When a student chooses the right alternative of an MC item, then he/she could get 3 points.

2.4.2. *Example of CR items.* CR item 1: Students are asked to use the experiment apparatus to make oxygen and verify its production. Students use the mouse to drag and drop the apparatus into the workspace, click "OK" when completed, and click "redo" to start again. The time limit for answering the CR item 1 is 5 minutes. The interface for CR item 1 is shown in Figure 2 and the problem solving strategy analysis of it is shown in Figure 13.

The definitions of decision nodes of CR item 1 are the followings:

① If the problem solving process is blank, it would be encoded 99 and scored 0.

② If the problem solving process appears "use conical flask + (hydrogen peroxide + carrot) + funnel + glass plate + incense", then it would be encoded 0 and scored 10 because of the correct procedural response (answer).

③ If the problem solving process appears "use conical flask", then it is considered as "can use conical flask in experiment", if not, then it means "don't know how to use conical flask in experiment", encoded 20.

TABLE 1. Representations of computerized response codes in CR item 1

BO1_: drag a conical flask into the workspace (Step ①)·····(see Figure 4)
LO1_: drag a funnel into the workspace and place it on top of the conical flask (Step ②)·····(see Figure 5)
LO4_: drag a funnel into the workspace but not place it on top of the conical flask (Step ④)·····(see Figure 6)
LO5_: drag a funnel into the workspace but is stuck and then return it to equipment apparatus
LO9_: drag a funnel outside the workspace
LE1_: drag an incense into the workspace and place it on top of the conical flask (Step ⑧)·····(see Figure 7)
LE4_: drag an incense into the workspace but not place it on top of the conical flask
LE5_: drag an incense into the workspace but is stuck and then return it back to equipment apparatus
LE9_: drag a incense outside the workspace
GL1_: drag a glass plate into the workspace and place it on the conical flask (Step ⑥)·····(see Figure 8)
GL4_: drag a glass plate into the workspace but not place it on top of the conical flask (Step ⑦)·····(see Figure 9)
GL5_: drag a glass plate into the workspace but is stuck and return it back to the area of equipment apparatus
GL9_: drag a glass plate outside the workspace
SU1_: pour baking soda into the conical flask through a funnel
SU2_: pour baking soda into the conical flask directly
SU4_: pour baking soda into the workspace but not into the flask
SU5_: pour baking soda into workspace but is stuck and return it back to the area of equipment apparatus
SU9_: pour baking soda outside the workspace
HU1_: pour diced carrots into the flask through funnel
HU2_: pour diced carrots into the conical flask directly (Step ⑤)·····(see Figure 10)
HU4_: drag diced carrots into the workspace and but not place them into the conical flask
HU5_: drag diced carrots into the workspace but is stuck and returned to equipment
HU9_: drag diced carrots moved outside the workspace
OO1_: pour hydrogen peroxide through the funnel into the conical flask (Step ③)·····(see Figure 11)
OO2_: pour peroxide directly into the conical flask (will spill)
OO4_: pour peroxide into the workspace but not into the conical flask
OO5_: pour peroxide into the conical flask but is stuck and return it back to the area of equipment apparatus
OO9_: drag peroxide outside the workspace
CH1_: pour vinegar into the conical flask through a funnel
CH2_: pour vinegar directly into the conical flask (will spill)
CH4_: pour vinegar into the workspace but not into the conical flask
CH5_: pour vinegar into the conical flask but is stuck and return it back to the area of equipment apparatus
CH9_: pour vinegar outside the workspace
CA1_: pour clarified limewater into flask through funnel
CA2_: pour clarified limewater directly into the flask (will spill)
CA4_: pour clarified limewater into the workspace but not into the conical flask
CA5_: pour clarified limewater into the flask but is stuck and return it back to the area of equipment apparatus
CA9_: pour clarified limewater outside the workspace
add_: press the "redo" button
right: Press the "OK" button (Step ⑨)·····(see Figure 12)

TABLE 2. The checklist for MC item 1

Sub-skill (s1)	1-1-1 know the phenomenon of combustion			
Question	1. ( ) When we light the candles if the electric power does not work. Which one is the right description? (1) Candle is a comburent. (2) When candles burn, there is no other substance, so candle will burn out and disappear completely. (3) When candle burns, it will be lighter around it. (4) 1kg of candle will turn to liquid when it burns, and it will turn into 1kg of liquid candle finally.			
Alternatives	alternative 1	alternative 2	alternative 3	alternative 4
Misconception	Unclear concept on comburent	Unclear about the product in burning	◎	Unclear about the product in burning
code	b5	b16		b16

④ If the problem solving process appears "hydrogen peroxide + carrot", then it means "has the competence to make oxygen with hydrogen peroxide and carrot". If not, then it means "does not have the competence to make oxygen".

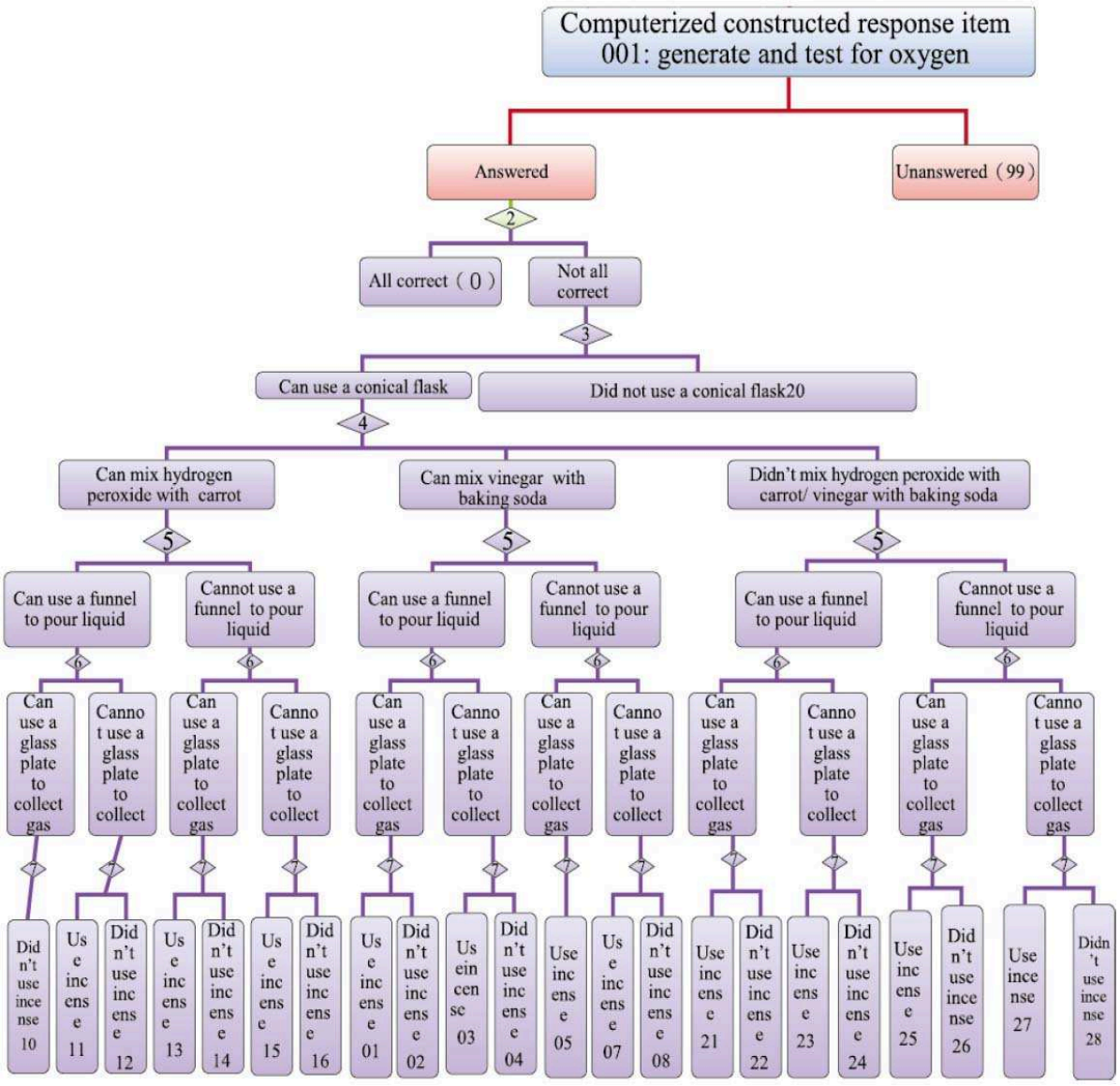

FIGURE 13. Problem solving strategy analysis of CR item 1

⑤ If the problem solving process appears "funnel", then it means "has the competence of using funnel and pour liquid into the conical flask", if not, then it means "does not have the competence of using a funnel".

⑥ If the problem solving process appears "glass plate", then it means "has the competence to collect oxygen from conical flask", if not, then it means "does not have the competence to use glass plate to collect oxygen".

⑦ If the problem solving process appears "incense", it means "has the competence to verify oxygen", if not, then it means "does not have the competence to verify oxygen".

For further computerized analysis, we defined problem solving strategy into word string listed in Table 1. The problem solving strategy word strings are based on the problem solving strategy flowchart in Figure 13.

For example, M_0 represents the correct problem solving strategy, which contains 6 different answering word string patterns representing the same problem solving strategy of M and gets the score of 10 points.

Another example is the problem solving strategy coded M_{27}, based on the problem solving strategy flowchart in Figure 13, pointing out that this student has the competence to use a conical flask, but lacks the competence to make oxygen (hydrogen peroxide + carrots), is unable to pour liquid with a funnel, unable to use glass plate to collect gas, and has the competence to use incense to verify oxygen. According to our analysis, there are 28 different answering word string patterns representing the same problem solving strategy of M_{27} whose score is 4 points.

The number of different answering word string patterns is the sum of students' different word strings representing the same problem solving strategy for CR item 1. For example, M_0 has 6 different problem solving strategies, and M_{11} has a total of 12 different problem solving strategies. Owing to layout limitation, we only list some examples of different word strings (see Table 3).

TABLE 3. Problem solving strategies of CR item 1

Code of problem solving strategy	Different word strings	Number of different answering word string patterns representing the same problem solving strategy
M_0	BO1_HU2_HU2_HU2_OO1_OO1_GL1_LE1_	6
	BO1_HU2_HU2_OO1_GL1_LE1_	
	BO1_HU2_OO1_GL1_LE1_	
	BO1_OO1_HU2_GL1_LE1_	
	.....	
M_{11}	BO1_GL1_OO1_HU2_LE1_	12
	BO1_HU2_HU2_OO1_LE1_	
	BO1_HU2_OO1_LE1_	
	BO1_HU2_OO1_LE1_GL1_	
	BO1_SU2_CH2_	
	.....	
M_{27}	BO1_HU2_HU2_OO2_SU2_LE1_ GL1_	28
	BO1__OO2_OO2_HU2_HU2_SU2_LE1	
	.....	

Due to the purpose of computerized diagnostic inference, the analytic and scoring process is proceeded by computer. Therefore, we have to transfer students' answering procedures on the screen of computer into different word strings like BO1_HU2_HU2_HU2_OO1 _OO1_GL1_LE1_ or BO1_HU2_HU2_OO1_GL1_LE1_, in order to be recorded into the database of the computerized diagnostic test system.

When a student finishes CR item 1, then he/she could get different scores (non-dichotomous) which range from 0 to 10 points according to his/her problem solving procedures.

As to the scoring rule (logic) of CR item 1 is that if the student lacks one certain competence which leads to construct one or some certain misconceptions, 2 points would be subtracted from the total score of 10. The score rubric related to different problem solving strategies of CR item 1 is listed in Table 4.

The problem solving strategies and scoring rules of CR item 2 and item 3 follow the similar logic and pattern of CR item 1. Therefore, they will not be described again in this paper. As far as all 3 CR items are concerned, the total sore of CR item 1 is 10 points, the total sore of CR item 2 is 12 points and the total sore of CR item 3 is 9 points. Combined with 69 points from 23 MC items, the total score of this test is 100 points.

TABLE 4. Score rubric of CR item 1

Code of problem solving strategy	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Score	10	8	6	6	4	6	4	4	2
Code of problem solving strategy	M_{10}	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}	M_{20}	M_{21}
Score	8	8	6	8	6	6	4	0	8
Code of problem solving strategy	M_{22}	M_{23}	M_{24}	M_{25}	M_{26}	M_{27}	M_{28}	M_{99}	M_{90}
Score	6	6	4	6	4	4	2	0	0

2.5. **The interface design of CR items.** In the study, the FLASH software is used not only to create an animated CR item interface, but also to collect the patterns of answering responses corresponding to different problem solving strategies the examinees used. The recordings of position coordinates, moving and collisions of objects in workspace designed by Flash software help us to determine the types of problem solving strategies and score them.

2.6. **Accuracy performance.** We train the testing data and evaluate students' misconceptions and sub-skills based on 5 fold cross validation. For evaluating the accuracy of aforementioned BN diagnostic mechanism, some criterions are described as follows: Accuracy (the correct classification rate of students' misconceptions and sub-skills) rate: The computing symbols related to accuracy are listed in Table 5 in which f_{11} denotes Experts' Criterion is yes (1) and the BN inference is also yes (1), f_{10} denotes Experts' Criterion is yes (1) and the BN inference is no (0), f_{00} denotes Experts' Criterion is no (0) and the BN inference is neither no (0), and f_{01} denotes Experts' Criterion is no (0) and the BN inference is yes (1).

TABLE 5. The formulation of accuracy

Experts' criterion \ BN inference	Yes (1)	No (0)
Yes (1)	f_{11}	f_{10}
No (0)	f_{01}	f_{00}

The definition of accuracy (in which N is the number of testing samples) in this paper is $\frac{f_{11}+f_{00}}{N}$.

3. Results.

3.1. **Common misconceptions and sub-skills.** According to the literature, teaching guides and discussions of research teams, students' common misconceptions and sub-skills for 5^{th} grade students in science and life technology learning domain are listed in Table 6 and Table 7.

3.2. **Diagnostic model based on BN.** The BN diagnostic model (See Figure 14) is composed of nodes which are 26 test items (23 MC items and 3 CR items), students' 27 misconceptions and 22 sub-skills. The test item is denoted as "I" (MC items) and CRI (CR items). The misconception is denoted as "b". The sub-skill is denoted as "s".

TABLE 6. Summary of misconceptions

1-1 Students know combustion requires oxygen	b1 unclear that combustion requires oxygen
	b2 unclear that oxygen is odorless and colorless
	b3 unclear that oxygen is everywhere
	b4 unclear about the concept of "burning point"
	b5 unclear about the concept of "comburent"
	b6 It is taken for granted that metals will not burn
	b7 unclear that the concept of "combustibles"
	b11 unclear about the 3 criteria of combustion
	b12 mix the concepts of combustibles and comburent
	b16 unclear about the product in burning
1-3 Students can make and verify carbon dioxide	b8 unclear that CO_2 cannot help combustion
	b9 unclear why CO_2 turns clear limewater murky
	b10 unclear that CO_2 is colorless and odorless
	b17 unclear that soda powder and vinegar can make CO_2
	b18 unclear about the application of CO_2
	b26 unclear how to collect CO_2 by pressing the plastic bag
1-2 Students can make and verify oxygen	b13 unclear that diced carrots and peroxide can create oxygen
	b14 cannot verify oxygen using experiment equipment
	b15 cannot apply oxygen in daily life
	b24 cannot use the funnel for pouring liquid into the conical flask
	b25 unclear how to use the conical flask
2-2 Students have the competence to make a simple fire extinguisher	b19 unclear how to use a fire extinguisher
	b20 unclear what materials should be used to extinguish a fire
	b27 cannot separate two items by using apparatus
2-3 Students understand the principles and methods of fire prevention and treatment	b21 unclear about the classification and function of fire extinguishers
	b22 unclear if metals are heated, they will have a high temperature
	b23 unclear to crouch for breathing fresh air in case of a fire

TABLE 7. Summary of sub-skills

Unit goals	Sub-skills
1-1 Students know combustion requires oxygen	s1 Students know the phenomenon of combustion
	s2 Students detect combustion requires oxygen
1-2 Students can make and verify oxygen	s3 Students have the competence to make oxygen
	s4 Students have the competence to verify oxygen
	s5 Students know the characteristics of oxygen
	s6 Students know the applications of oxygen in daily life
1-3 Students can make and verify carbon dioxide	s7 Students can make carbon dioxide
	s8 Students can detect CO_2 does not help with combustion
	s9 Students can detect CO_2 will turn clarified limewater murky
	s10 Students can say the characteristics of CO_2
	s11 Students can verify the bubbles in the soda is CO_2
	s12 Students can say the application of CO_2 in daily life
2-1 Students have the competence to extinguish a fire	s13 Students understand the meaning of combustibles
	s14 Students understand the meaning of comburent
	s15 Students understand the meaning of burning point
	s16 Students know the 3 criteria for combustion: combustible, comburent and burning point.
	s17 Students understand the principle and means of extinguish a fire
2-2 Students have the competence to make a simple fire extinguisher	s18 Students can design and make a simple fire extinguisher
	s22 Students can explain the method of using fire extinguisher
2-3 Students understand the principles and methods of fire prevention and treatment	s19 Students know ways of fire prevention
	s20 Students can explain the procedures to take in case of a fire
	s21 Students can explain how to escape in case of a fire

In this model, the statistical model used BN mechanism to update misconception variables and sub-skill variables with conditional probabilities by collecting students' answering responses from both MC items and CR items. Basically, a conditional probability gives an estimation for the likelihood that student is at a certain level of interested variables given all relevant data collected so far.

3.3. **Evaluation of the accuracy of diagnostic performance.** By comparing the classifications of the problem solving strategies judged by experts and those judged by computer. The accuracy rates of CR item 1, CR item 2 and CR item 3 are 94.45%, 96% and 97.92% (see Table 8), with an average accuracy of 96.13%.

Table 8. Accuracy rates of CR items

Item no.	Accuracy
CR item 1	94.45%
CR item 2	96%
CR item 3	97.92%
Average	96.13%

The overall accuracy rates of CR items shows quite accurate diagnostic results for CR items by computer. Based on accurate diagnostic results for CR items, the corresponding automated scoring of CR items can lead to good diagnostic accuracy rates of misconceptions and sub-skills of students.

The accuracy rates (comparison between computerized diagnosis and experts' diagnosis) of misconceptions and sub-skills of this test combined with 23 MC items and 3 CR items based on BN are listed in Table 9 and Table 10.

Table 9. Accuracy rate of misconceptions

No. of misconception	b01	b02	b03	b04	b05	b06	b07	b08
Accuracy rate (%)	96.36	93.76	95.67	91.86	80.08	88.22	94.11	89.08
No. of misconception	b09	b10	b11	b12	b13	b14	b15	b16
Accuracy rate (%)	88.21	94.1	100	89.78	66.21	82.16	94.1	83.37
No. of misconception	b17	b18	b19	b20	b21	b22	b23	b24
Accuracy rate (%)	73.83	87.36	99.66	93.94	100	94.27	98.09	88.39
No. of misconception	b25	b26	b27					
Accuracy rate (%)	99.31	74.52	94.11					

Average of accuracy rates of misconceptions: 90.02%

Table 10. Accuracy rate of sub-skills

No. of sub-skills	s01	s02	s03	s04	s05	s06	s07	s08
Accuracy rate (%)	95.32	91.86	97.05	82.85	83.37	93.24	71.41	92.89
No. of sub-skills	s09	s10	s11	s12	s13	s14	s15	s16
Accuracy rate (%)	85.79	89.95	84.07	78.86	92.89	92.2	95.49	100
No. of sub-skills	s17	s18	s19	s20	s21	s22		
Accuracy rate (%)	100	100	99.66	91.5	87.36	100		

Average of accuracy rates of sub-skill: 91.17%

There is an average accuracy rate of 90.02% on diagnosis of misconceptions (see Table 9), and there is an average accuracy rate of 91.17% on diagnosis of sub-skills (see Table 10). One example report of the student's diagnostic result is shown in Figure 15.

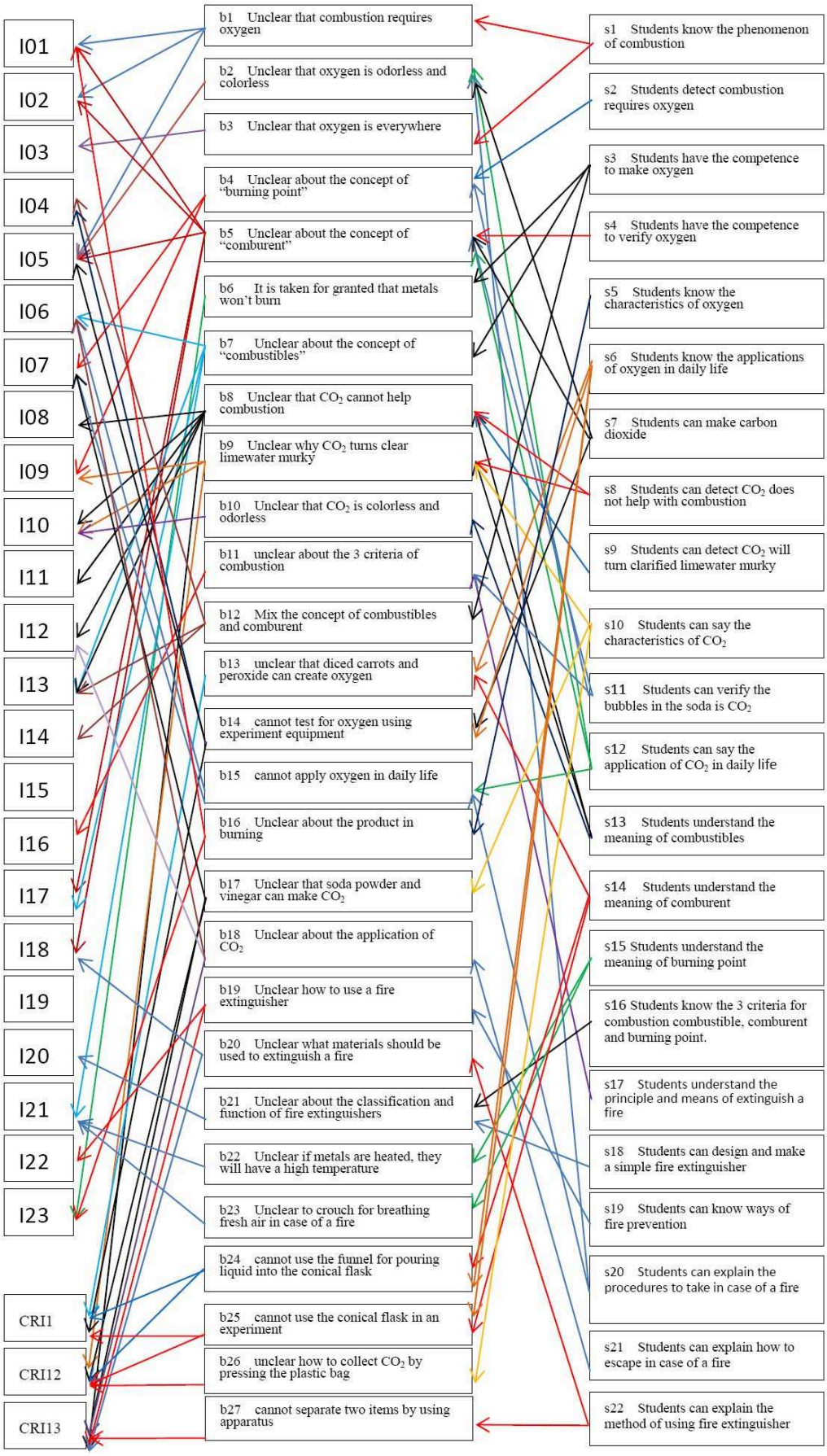

FIGURE 14. Bayesian network diagram

錯誤類型	診斷結果
b1 不清楚氧氣可以幫助物質燃燒	O
b2 不了解氧氣具有無色無味的特性	O
b3 不了解空氣到處都有的概念	O
b4 燃點這個名詞觀念不清	O
b5 助燃物這個名詞觀念不清	X
b6 誤以為金屬不可能燃燒	X
b7 可燃物這個名詞觀念不清	X
b8 不了解二氧化碳不能幫助燃燒	X
b9 不了解二氧化碳會使澄清石灰水變混濁	O
b10 不了解二氧化碳具有無色無味的特性	O
b11 不知道燃燒的三要素	O
b12 將可燃物與助燃物的概念混淆	O
b13 不知道胡蘿蔔丁加雙氧水可產生氧氣	X
b14 不知道胡蘿蔔丁加雙氧水可產生氧氣	X
b15 無法在日常生活中應用氧氣	X
技能與概念	**診斷結果**
s13 能認識可燃物的意義	X
s14 能認識助燃物的意義	X
s15 能認識燃點這個名詞的意義	X
s16 能認識燃燒需要三個條件：可燃物、助燃物、達到燃點	X
s17 能了解滅火的原理及方式	X
s18 能學習設計及製作簡易滅火器	X
s22 能說明滅火器的使用方法	X
s19 能認識火災的預防方法	X
s20 能說明火災的處理方法	X
s21 能說明火災發生時的逃生方法	X

FIGURE 15. Student's diagnostic test report

4. **Discussion and Conclusions.** In this work, we have presented a good use of BN algorithm in diagnosing students' misconceptions and sub-skills by computer instead of manpower. Through our new BN diagnostic model integrating students' misconceptions, sub-skills, MC items and CR items, the computerized test system could diagnose students' conceptual knowledge and procedural knowledge immediately and accurately. In the BN diagnostic model, nodes which represent students' misconceptions and sub-skills of science knowledge have a well-defined semantics and links between them. Therefore, educators or researchers could accurately describe the educational relationships between them.

The validity of the proposed diagnostic model has been tested by comparing the diagnostic results of computer and those of experts. The results (the overall average of immediate diagnostic accuracy rates of misconceptions and sub-skills are both above 90%) obtained are very promising. In other words, the diagnostic test and analytic design in this study can have highly accurate estimations of the student's cognitive states. The present case study in the "Air and Combustion" unit has illustrated the good use of BN mechanism in diagnosing students' misconceptions and sub-skills by computer instead of manpower. In addition to the traditional MC items, students' misconceptions and sub-skills based on procedural knowledge can also be detected by CR items.

Moreover, the additional and influential misconceptions (b24, b25, b26 and b27) can be found (detected) from CR items. Through CR items, we could detect students' misconceptions which could not be detected by MC items. The additional misconceptions detected through CR items are listed in Table 11. As we described in the previous introduction section, students really need to have a more comprehensive understanding in the contents of science learning in order to construct their own response to the CR items based on that understanding. Based on the demand and experience for answering CR items, students may develop abilities to comprehend factual science knowledge, to synthesize ideas into an explanation, to use evidence to support ideas and to analyze a graph or diagram. In many instances, these aforementioned abilities could contribute to their valid science inferences and learning. These findings could provide very precious and useful information for educators and researchers.

TABLE 11. Additional misconceptions detected through CR items

b24	cannot use the funnel for pouring liquid into the conical flask
b25	unclear how to use the conical flask
b26	unclear how to collect CO_2 by pressing the plastic bag
b27	cannot separate two items by using apparatus

However, even though the results obtained are very satisfactory, it has been possible to improve the authenticity of the test, if the experimental equipment in the interference was presented by real pictures. Besides, we must insist again that, in spite of the excellent results, this empirical evaluation should be only considered and defined as a successful case study. Furthermore, before BN could be used in a comprehensive analysis of educational tests, the correct test development, test implementation, and decision making (BN structure, threshold definition, judgments of students' latent trait states) must be well defined and examined by experts (teachers and professors) thoroughly.

We lay emphasis on assessing students' operation, analysis, integration and thinking abilities by computerized test instead of using lots of manpower. In general, BN has a flexible structure, fitting to a wide range of students' answering response patterns. Its ability for abductive reasoning and uncertainty handling makes it a suitable technique for computerized evaluation of students' hidden variables (misconceptions and sub-skills) for educational purposes. Since students' answering responses were well recorded and analyzed, we could know the strengths and weaknesses of the students' learning situations and the types of misconceptions they developed. Therefore, appropriate remedial learning materials could be provided automatically and properly by computer.

In contrast to other classical test theories or famous modern test theories like item response theory (IRT), BN in this paper updates the prior probabilities by propagation of new observations through the related network, yielding posterior probabilities. These posteriors, unlike priors that are based mainly on generic data and expert knowledge, are

more specific to the evaluation of students' latent traits and better reflect its characteristics probabilistically in real situations according to the points of views. Therefore, BN is a very useful algorithm in situations where there is not enough information and manpower to estimate the related values of interested variables (misconception and sub-skills in this study) for educational purposes and real situations.

From above, the major contributions of the present study are:

i) This paper proposes a new automated diagnostic algorithm for CR items based on BN.
ii) This paper proposes a new automated scoring mechanism for considering MC items and CR items together.
iii) This paper proposes a computerized diagnostic test design for assessing examinees' conceptual knowledge and procedural knowledge together.
iv) This paper develops a practical computerized diagnostic test system composed of MC items and CR items.
v) This paper examines the validities of the proposed algorithm and scoring mechanism.

Most importantly, this paper is, to our limited knowledge and besides our investigation, the rare one using empirical data to construct and apply the BN topology to both MC items and CR items in a computerized diagnostic test. Through the design and practice of this computerized diagnostic test system with CR items, we overcome the time-consuming problems for scoring the CR items and build good rubrics of them to minimize the subjectivities of graders. Besides, based on the proposed computerized diagnostic mechanism in this paper, the results show the fundamental and possibilities of practical applications in large-scale educational assessments like TIMSS, PIRLS or PISA.

Due to the great performances of automation and accuracy in the above-mentioned computerized diagnostic test, teachers and researchers could apply the design and analytic method in this study not only to diagnostic assessments but also to placement assessment, formative assessments and summative assessments in science classes. Moreover, the research methodology and results in this paper could be the academic foundation for cognitive curriculum specialists, psychologists (learning emphasis), teacher educators (Curriculum and Instruction emphasis), and measurement experts (assessment emphasis) for different purposes and applications.

Regarding future work, there are several directions to be explored, which are that we may add more samples (then the BN will have more responses to train with, leading to the increase of its reliability), we may develop a more advanced BN algorithm like Hierarchical Bayesian Networks for the diagnostic purposes, and we may develop more different types of items for various units in science curriculums.

In recent years, researchers have made a great deal of progress in using computers to score the responses to items. Automated scoring offers the possibility of greatly reducing the time and cost of the scoring process, making it more practical to use constructed response items in real testing situations where human scoring would be impractical or prohibitively expensive.

Acknowledgment. This research is partially supported by the "Ministry of Science and Technology, Taiwan" under Grant No. NSC 97-2511-S-142-004- and Grant No. MOST 103-2511-S-142-010 -MY3.

REFERENCES

[1] A. Maerlender, L. Flashman, A. Kessler et al., Examination of the construct validity of ImPACT^{TM} computerized test, traditional, and experimental neuropsychological measures, *The Clinical Neuropsychologist*, vol.24, no.8, pp.1309-1325, 2010.

[2] D. J. Ma, H. K. Yang and J.-M. Hwang, Reliability and validity of an automated computerized visual acuity and stereoacuity test in children using an interactive video game, *American Journal of Ophthalmology*, vol.156, no.1, pp.195-201, 2013.
[3] M. E. Poehner and J. P. Lantolf, Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA), *Language Teaching Research*, vol.17, no.3, pp.323-342, 2013.
[4] R. C. Moore, A. L. Harmell, J. Ho et al., Initial validation of a computerized version of the UCSD Performance-Based Skills Assessment (C-UPSA) for assessing functioning in schizophrenia, *Schizophrenia Research*, vol.144, no.1, pp.87-92, 2013.
[5] J. Liu, Z. Ying and S. Zhang, A rate function approach to computerized adaptive testing for cognitive diagnosis, *Psychometrika*, pp.1-23, 2013.
[6] X. Mao and T. Xin, The application of the Monte Carlo approach to cognitive diagnostic computerized adaptive testing with content constraints, *Applied Psychological Measurement*, vol.37, no.6, pp.482-496, 2013.
[7] S. Brown, J. Bull and P. Race, *Computer-Assisted Assessment of Students: Routledge*, 2013.
[8] H.-C. Chu and S.-C. Chang, Developing an educational computer game for migratory bird identification based on a two-tier test approach, *Educational Technology Research and Development*, pp.1-15, 2013.
[9] Y.-L. Chen, P.-R. Pan, Y.-T. Sung et al., Correcting misconceptions on electronics: Effects of a simulation-based learning environment backed by a conceptual change model, *Educational Technology & Society*, vol.16, no.2, pp.212-227, 2013.
[10] C. Nixon and P. E. Kennedy, Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions, *Southern Economic Journal*, vol.68, no.4, pp.957-971, 2002.
[11] W. C. Ward and R. E. Bennett, *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment: Routledge*, 2012.
[12] C. R. Reynolds, R. B. Livingston, V. L. Willson et al., *Measurement and Assessment in Education: Pearson Education International*, 2010.
[13] H. Wainer and D. Thissen, Combining multiple-choice and constructed-response test scores: Toward a marxist theory of test construction, *Applied Measurement in Education*, vol.6, no.2, pp.103-118, 1993.
[14] B. Benson, J. R. Bergan, S. Cunningham et al., Item banking system for standards-based assessment, *Google Patents*, 2014.
[15] J. W. Pellegrino, Proficiency in science: Assessment challenges and opportunities, *Science*, vol.340, no.6130, pp.320-323, 2013.
[16] S. A. Hill, National assessment of educational progress, *American Mathematical Monthly*, vol.87, no.6, pp.427-428, 1980.
[17] I. V. Mullis, M. O. Martin, G. J. Ruddock et al., TIMSS 2011 assessment frameworks, *International Association for the Evaluation of Educational Achievement*, 2009.
[18] C. H. Tienken, TIMSS implications for US education, *Science*, vol.7, 2013.
[19] E. A. Hanushek and L. Woessmann, The role of international assessments of cognitive skills in the analysis of growth and development, *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*, pp.47-65, 2013.
[20] I. Kirsch, M. Lennon, M. von Davier et al., On the growing importance of international large-scale assessments, *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*, pp.1-11, 2013.
[21] S. Thomson, L. de Bortoli, M. Nicholas et al., *Highlights from the Full Australian Report: Challenges for Australian Education: Results from PISA 2009*, 2013.
[22] J. R. Raker, J. M. Trate, T. A. Holme et al., *Adaptation of an Instrument for Measuring the Cognitive Complexity of Organic Chemistry Exam Items*, 2013.
[23] S. Kim and T. Moses, Determining when single scoring for constructed-response items is as effective as double scoring in mixed-format licensure tests, *International Journal of Testing*, vol.13, no.4, 2013.
[24] J. W. Pellegrino, Proficiency in science: Assessment challenges and opportunities, *Science*, pp.320-323, 2013.
[25] K. Abida, *Assessing Students' Math Proficiency Using Multiple-Choice and Short Constructed Response Item Formats*, www.cg.publisher.com, 2011.

[26] M. C. Rodriguez, *Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations*, 2003.
[27] M. E. Martinez, *A Comparison of Multiple-Choice and Constructed Figural Response Items*, 1991.
[28] R. Lukhele, D. Thissen and H. Wainer, On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests, *Journal of Educational Measurement*, vol.31, no.3, pp.234-250, 1994.
[29] P. Harik, P. Baldwin and B. Clauser, Comparison of automated scoring methods for a computerized performance assessment of clinical judgment, *Applied Psychological Measurement*, vol.37, no.8, 2013.
[30] E. A. Sheaffer and R. T. Addo, Pharmacy student performance on constructed-response versus selected-response calculations questions, *American Journal of Pharmaceutical Education*, vol.77, no.1, 2013.
[31] M. Kastner and B. Stangla, Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia – Social and Behavioral Sciences*, vol.12, pp.263-273, 2011.
[32] O. P. F. I. S. Assessment, *PISA Computer-Based Assessment of Student Skills in Science*, Paris and Washington D.C., Organisation for Economic Co-operation and Development, 2010.
[33] C.-Y. Kuo and H.-K. Wu, Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science, *Computers & Education*, vol.68, pp.388-403, 2013.
[34] M. Jakubowski, *Analysis of the Predictive Power of PISA Test Items*, Organisation for Economic Cooperation and Development (OECD), 2013.
[35] H.-K. Wu, S. W.-Y. Lee, H.-Y. Chang et al., Current status, opportunities and challenges of augmented reality in education, *Computers & Education*, 2012.
[36] Y. Zhao, F. Xiao and S. Wang, An intelligent chiller fault detection and diagnosis methodology using Bayesian belief network, *Energy and Buildings*, vol.57, pp.278-288, 2013.
[37] A. Oniśko and M. J. Druzdzel, Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems, *Artificial Intelligence in Medicine*, vol.57, no.3, pp.197-206, 2013.
[38] K. S. Kim and Y. S. Choi, Computerized adaptive testing and learning using Bayesian network, *Proc. of the Companion Publication of the 2013 International Conference on Intelligent User Interfaces Companion*, pp.91-92, 2013.
[39] G. Corani, C. Magli, A. Giusti et al., A Bayesian network model for predicting pregnancy after in vitro fertilization, *Computers in Biology and Medicine*, vol.43, no.11, pp.1783-1792, 2013.
[40] C.-Y. Ting, Y.-C. Sam and C.-O. Wong, Model of conceptual change for INQPRO: A Bayesian network approach, *Computers & Education*, 2013.
[41] D. Heckerman and M. P. Wellman, Bayesian networks, *Communications of the ACM*, vol.38, no.3, pp.27-30, 1995.
[42] J. Lee and J. E. Corter, Diagnosis of subtraction bugs using bayesian networks, *Applied Psychological Measurement*, vol.35, no.1, pp.27-47, 2011.
[43] G. Castillo, L. Descalço, S. Diogo et al., Computerized evaluation and diagnosis of student's knowledge based on Bayesian networks, *Sustaining TEL: From Innovation to Learning and Practice, Lecture Notes in Computer Science*, pp.494-499, 2010.
[44] A. Zagorecki, P. Orzechowski and K. Hołownia, Online diagnostic system based on Bayesian networks, *Artificial Intelligence in Medicine, Lecture Notes in Computer Science*, pp.145-149, 2013.
[45] P. García, A. Amandi, S. Schiaffino et al., Evaluating Bayesian networks' precision for detecting students' learning styles, *Computers & Education*, vol.49, no.3, pp.794-808, 2007.
[46] J. Vomlel, Bayesian networks in educational testing, *International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems*, vol.12, pp.83-100, 2004.
[47] V. Jiří, Bayesian networks in educational testing, *Proc. of the 1st European Workshop on Probabilistic Graphical Models*, pp.176-185, 2002.
[48] J. Vomlel, Building adaptive tests using Bayesian networks, *Kybernetika*, vol.40, no.3, pp.333-348, 2004.
[49] S.-C. Shih and B.-C. Kuo, Using Bayesian networks for modeling students' learning bugs and subskills, *The 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, pp.69-75, 2005.
[50] H. Modell, J. Michael and M. P. Wenderoth, Helping the learner to learn: The role of uncovering misconceptions, *The American Biology Teacher*, vol.67, no.1, pp.20-26, 2005.
[51] M.-H. Chiu and R. Duit, *Globalization: Science Education from an International Perspective*, 2011.

[52] R. Duit, D. F. Treagust and H. Mansfield, Investigating student understanding as a prerequisite to improving teaching and learning in science and mathematics, *Improving Teaching and Learning in Science and Mathematics*, pp.17-31, 1996.
[53] M. H. Schneps, J. Ruel, G. Sonnert et al., Conceptualizing astronomical scale: Virtual simulations on handheld tablet computers reverse misconceptions, *Computers & Education*, 2013.
[54] A. C. Maskiewicz and J. E. Lineback, Misconceptions are "So Yesterday!", *CBE-Life Sciences Education*, vol.12, no.3, pp.352-356, 2013.
[55] W. Stout, Skills diagnosis using IRT-based continuous latent trait models, *Journal of Educational Measurement*, vol.44, no.4, pp.313-324, 2007.
[56] C.-Q. Lee and H.-C. She, Facilitating students' conceptual change and scientific reasoning involving the unit of combustion, *Research in Science Education*, vol.40, no.4, pp.479-504, 2010.
[57] K. Salta and C. Tzougraki, Conceptual versus algorithmic problem-solving: Focusing on problems dealing with conservation of matter in chemistry, *Research in Science Education*, vol.41, no.4, pp.587-609, 2011.
[58] A. Thanukos and J. Scotchmoor, Making connections, *Evolution Challenges: Integrating Research and Practice in Teaching and Learning about Evolution*, p.410, 2012.
[59] O. L. Liu, C. Brew, J. Blackmore et al., Automated scoring of constructed-response science items: Prospects and obstacles, *Educational Measurement: Issues and Practice*, 2014.
[60] R. G. Almond, Using automated essay scores as an anchor when equating constructed response writing tests, *International Journal of Testing*, vol.14, no.1, pp.73-91, 2014.
[61] K. S. Taber, Meeting educational objectives in the affective and cognitive domains: Personal and social constructivist perspectives on enjoyment, motivation and learning chemistry, *Affective Dimensions in Chemistry Education*, pp.3-27, 2015.
[62] B. Cope and M. Kalantzis, *The Powers of Literacy (RLE Edu I): A Genre Approach to Teaching Writing: Routledge*, 2014.
[63] H.-C. Chu and C.-M. Hung, Effects of the digital game-development approach on elementary school students' learning motivation, problem solving, and learning achievement, *International Journal of Distance Education Technologies*, vol.13, no.1, pp.87-102, 2015.
[64] T. Shimoda, B. White, M. Borge et al., Designing for science learning and collaborative discourse, *Proc. of the 12th International Conference on Interaction Design and Children*, pp.247-256, 2013.
[65] Y. J. Dori and Z. Kaberman, Assessing high school chemistry students' modeling sub-skills in a computerized molecular modeling learning environment, *Instructional Science*, vol.40, no.1, pp.69-91, 2012.
[66] S. DiGangi, J. Gorin, C. H. Yu and A. Jannasch-Pennell, *A Multi-Disciplinary Approach to Cognitive-Based Assessment*, Seattle, WA, 2007.
[67] V. Shute and J. Underwood, Diagnostic assessment in mathematics problem solving, *Technology Instruction Cognition and Learning*, vol.3, nos.1/2, p.151, 2006.
[68] B. du Boulay, Computers and teacher education, in *World Yearbook of Education 1982/3: Computers and Education*, J. Megarry, D. R. F. Walker and S. Nisbet (eds.), p.177, 2013.
[69] L. Haapasalo and P. Samuels, Responding to the challenges of instrumental orchestration through physical and virtual robotics, *Computers & Education*, vol.57, no.2, pp.1484-1492, 2011.
[70] J. Hiebert, *Conceptual and Procedural Knowledge: The Case of Mathematics: Routledge*, 2013.
[71] M. Schneider, B. Rittle-Johnson and J. R. Star, Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge, *Developmental Psychology*, vol.47, no.6, p.1525, 2011.
[72] H.-C. Li, A comparative analysis of British and Taiwanese students' conceptual and procedural knowledge of fraction addition, *International Journal of Mathematical Education in Science and Technology*, pp.1-12, 2014.
[73] J. Leppavirta, H. Kettunen and A. Sihvola, Complex problem exercises in developing engineering students' conceptual and procedural knowledge of electromagnetics, *IEEE Trans. Education*, vol.54, no.1, pp.63-66, 2011.