Computers & Education 55 (2010) 527–540

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

A multi-expert approach for developing testing and diagnostic systems based on the concept-effect model

Patcharin Panjaburee [a], Gwo-Jen Hwang [b], Wannapong Triampo [c,*], Bo-Ying Shih [b]

[a] Institute for Innovative Learning, Mahidol University, 999, Phuttamonthon 4 Road, Salaya, Nakorn Pathom 73170, Thailand
[b] Department of Information and Learning Technology, National University of Tainan 33, Sec. 2, Shulin St., Tainan city 70005, ROC, Taiwan
[c] Department of Physics, Faculty of Science, Mahidol University, Rama VI, Bangkok 10400, Thailand

ARTICLE INFO

Article history:
Received 24 August 2009
Received in revised form
11 February 2010
Accepted 11 February 2010

Keywords:
Computer-based testing
Computer-assisted learning
Concept-effect relationships
Multiple expert systems

ABSTRACT

With the popularization of computer and communication technologies, researchers have attempted to develop computer-assisted testing and diagnostic systems to help students improve their learning performance on the Internet. In developing a diagnostic system for detecting students' learning problems, it is difficult for individual teachers to address the exact relationships between the test items and the concepts. To cope with this problem, this study proposes an innovative approach to eliciting and integrating the weightings of test item-concept relationships from multiple experts. Based on the proposed approach, a testing and diagnostic system has been implemented; moreover, an experiment was conducted to evaluate the performance of our approach. By analyzing the results from four groups of students using learning suggestions provided by different models, it was found that the learning performance of the students who received learning suggestions by applying the innovative approach was significantly better than for those who received guidance based on the original model.

1. Background and motivation

With the popularization of computers and communication technologies, information as well as instructional activities has been located on the Internet. Notable examples include the development of computer-assisted tutoring and testing systems (Chiou, Hwang, & Tseng, 2009; Hooper, 1992; Hwang, Cheng, Chu, Tseng, & Hwang, 2007; Hwang, Chu, Yin, & Lin, 2008; Hwang, Yang, Tsai, & Yang, 2009; Lee, Lee, & Leu, 2009; Plessis, Biljon, Tolmie, & Wollinger, 1995; Springer & Pear, 2008; Tsai & Chou, 2002; Tseng, Chu, Hwang, & Tsai, 2008). Specifically, during the tutoring process, online evaluation becomes an important form of learning feedback to provide the learning status of each student, and tests are mostly used as a way to gain this knowledge (Gronlund, 2003; Hwang et al., 2007; Tsai & Chou, 2002). In the past decades, researchers have shown the equivalence of paper-based and computer-based tests in terms of testing quality, implying that the development of computer-based testing systems and relevant techniques are worthwhile (Chiou et al., 2009; Hwang, 2003b; Hwang, Chu, et al., 2008).

Conventional testing systems represent the learning status of a student by assigning a total score or grade. Such feedback makes the students aware of their learning status through the score or grade; however, this information alone is insufficient for improving their learning performance unless further guidance can be given (Gerber, Grund, & Grote, 2008). This implies that providing learning suggestions for students after testing is an important research issue (Hwang, Tseng, & Hwang, 2008).

In recent years, researchers have proposed various approaches for developing adaptive learning systems based on the personal features or learning behaviors of students (Casamayor, Amandi, & Campo, 2009; Cheng, Lin, Chen, & Heh, 2005; Huang, Liu, Chu, & Cheng, 2007; Hwang, Tsai, Tsai, & Tseng, 2008; Manning & Dix, 2008; Offer & Bos, 2009). Furthermore, models or mechanisms for diagnosing student learning problems and providing personalized learning guidance have been presented as well (Bai & Chen, 2008; Huang, Lin, & Cheng, 2009; Hwang, Hsiao, & Tseng, 2003; Lee et al., 2009; Pavlekovic, Zekic-Susac, & Djurdjevic, 2009; Tseng, Sue, Su, Weng, & Tsai, 2007; Tung, Huang, Keh, & Wai, 2009). Among the existing models, the Concept-Effect Relationship (CER) model, which represents the prerequisite relationships among concepts in a course, has been proved to be an effective way of improving the learning performance of students (Hwang,

* Corresponding author. Tel.: +668 40042689; fax: +662 441 9322.
E-mail addresses: panjaburee_p@hotmail.com (P. Panjaburee), gjhwang@mail.nutn.edu.tw (G.-J. Hwang), wtriampo@gmail.com, scwtr@mahidol.ac.th (W. Triampo), shallmay14@gmail.com (B.-Y. Shih).

0360-1315/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2010.02.015

2003a). The CER model demonstrates a systematic procedure for identifying the learning problems of students for each concept taken into account. It has been used to successfully detect the learning problems of students and to give personalized suggestions to them for several science and mathematics courses (Hwang et al., 2007; Hwang, Tseng, et al., 2008; Jong, Lin, Wu, & Chan, 2004).

Although the CER model has shown its effectiveness in helping students improve their learning performance, past experiences of applying this model also reveal the difficulty of applying it. One of the major problems of applying the CER model is the need to define the weighting or degree of relevance for each concept to each test item. It is often the case that individual teachers might provide some imprecise <test item, concept> relationships owing to ignorance or subjectivity or unintentionally inconsistent decision making (Hwang, Tseng, et al., 2008; Lee et al., 2009); moreover, researchers have indicated that domain experts with different experiences could have different expertise or understanding of each portion of the knowledge; therefore, the cooperation of several experts (experienced teachers) has been suggested (Chu & Hwang, 2008; Huang & Shimizu, 2006; Hwang, Chen, Hwang, & Chu, 2006; Léger & Naud, 2009; Medsker, Tan, & Turban, 1995; Mittal & Dym, 1985). As the experts might have different experiences and domain knowledge or backgrounds, it becomes an important and challenging issue to integrate the opinions of multiple experts to obtain high quality <test item, concept> relationships such that more accurate and truthful learning suggestions can be given to the students (Chu & Hwang, 2008).

To cope with this problem, this paper presents an innovative approach by integrating <test item, concept> relationships from multiple experts. Moreover, a testing and diagnostic system based on this approach has been implemented, and an experiment was conducted to evaluate the performance of the innovative approach.

2. The concept-effect relationship (CER) model

Hwang (2003a) proposed the CER model to represent the prerequisite relationships among concepts that need to be learned in a dedicated order. Such a model has been referred to by several researchers in developing testing and diagnostic mechanisms or systems for improving the learning performance of students. Moreover, various applications have revealed the effectiveness of the CER model. For example, Jong, Chan, and Wu (2007) developed a learning behavior diagnosis system which was applied to a computer course of a university and yielded positive experimental results for both learning status and learning achievement. In the meantime, Tseng et al. (2007) employed the CER model to provide learning guidance for individual students in the physics course of a junior high school. Furthermore, Hwang, Tseng, et al. (2008) reported the effectiveness of the CER model in improving the learning achievements of students in a Mathematics course of an elementary school.

In the CER model, the diagnosis of student learning problems mainly depends on the prerequisite relationships between the concepts to be learned. Consider two concepts to be learned, say C_i and C_j. If C_i is a prerequisite to efficiently performing the more complex and higher level concept C_j, then a concept-effect relationship $C_i \rightarrow C_j$ is said to exist. For example, to learn the concept "subtraction of positive integer," one may first need to learn "addition of positive integer", while learning "division of positive integer" may require first learning "subtraction of positive integer" and "multiplication of positive integer". Fig. 1 presents an illustrative example of the concept-effect relationships, which are important in diagnosing student learning problems. For example, if a student fails to answer most of the test items concerning "division of positive integer" due to a lack of understanding of the questions posed or because of carelessness, the problem is likely because the student has not thoroughly learned "division of positive integer" or its prerequisite concepts (such as "subtraction of positive integer" or "multiplication of positive integer"). Therefore, teachers could identify the learning problems of students by tracing the concept-effect relationships (Cheng et al., 2005; Hwang, 2003a; Hwang, Tseng, et al., 2008).

In the CER model, all of the possible learning paths will be taken into consideration to find the poorly-learned paths. In the illustrative example given in Fig. 1, there are two learning paths for the subject unit:

PATH1 : $C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_5$

PATH2 : $C_1 \rightarrow C_2 \rightarrow C_4 \rightarrow C_5$

To provide learning suggestions to individual students, the error ratio (ER) for each student to answer the test items related to each concept needs to be analyzed; therefore, it is necessary to establish a Test Item Relationship Table (TIRT), which represents the degree of association between test item Q_i and concept C_k (Hwang, 2003a). An illustrative example of a TIRT comprising ten concepts and twelve test items is listed in Table 1, where the TIRT (Q_i, C_k) is a value ranging from 0 to 1; "1" represents "high relevance" and "0" represents "no

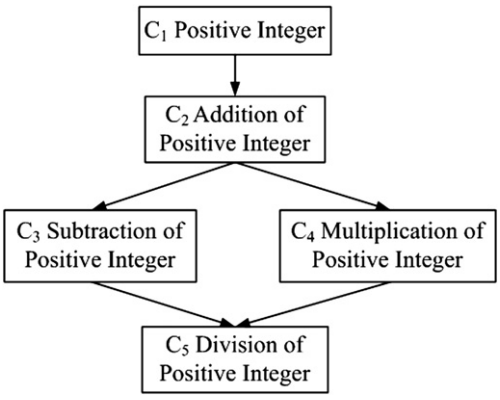

Fig. 1. Illustrative example of concept-effect relationships.

Table 1
Illustrative example of a Test Item Relationship Table (TIRT).

		Concept C_i									
		C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
Test item Q_t	Q_1	1	0	0	0.6	0	0	0.2	0	0	0
	Q_2	0	0.3	0	0	0	1	0	0	0	0
	Q_3	0.5	0	0	0	0.5	0	0	0	0	0
	Q_4	0	0.4	0.2	0	0	0	1	0.3	0.4	0
	Q_5	0	0	0.3	1	0	0	0.3	0	0	1
	Q_6	0.3	0.4	0	0.7	0.5	0	0	0	0	0
	Q_7	0	0	0	0	0	0.4	0	0	0.3	0
	Q_8	0.8	0	0	0	0	0	0.6	0	0	0
	Q_9	0.4	1	0	0	1	0.2	0	0.8	1	0.2
	Q_{10}	0	0	1	0	0	0	0	1	0	0
	Q_{11}	0	0	0	0	0	0.5	0	0	0	0.6
	Q_{12}	0	0	0	0.3	0	0	0	0	0	0

relevance". The error ratio for a student for concept C_i is then calculated by dividing the sum of TIRT (Q_i, C_k) values of the test items that the student failed to correctly answer by that of all of the test items.

Assuming that the error ratios (ER) for a student to answer the test items concerning C_1, C_2, C_3, C_4, and C_5 are 0.0, 0.25, 0.35, 0.2, and 0.6 respectively, we have

$$\text{PATH1}: C_1(0.0) \rightarrow C_2(0.25) \rightarrow C_3(0.35) \rightarrow C_5(0.6)\text{and}$$

$$\text{PATH2}: C_1(0.0) \rightarrow C_2(0.25) \rightarrow C_4(0.2) \rightarrow C_5(0.6)$$

A threshold θ is used to determine the acceptable error ratio. If $\text{ER}(C_j) \leq \theta$, the student is said to have learned concept C_j; otherwise, the student has failed to learn the concept, and it is selected as a node of the poorly-learned path. Assuming that the teacher has defined θ to be 0.3, the poorly-learned paths are as follows:

$$\text{PATH1}: C_3(0.35) \rightarrow C_5(0.6)\text{and}$$

$$\text{PATH2}: C_5(0.6)$$

Therefore, the learning problems of the student could be a misunderstanding of concepts C_3 and C_5; moreover, the student should learn C_3 before learning C_5.

Although several studies on CER-oriented models have demonstrated their benefits in providing learning advice to individual students, previous experiences of practical applications have also revealed some problems and difficulties of applying such models (Hwang et al., 2007; Hwang, Tseng, et al., 2008; Jong et al., 2004). For most science or mathematics courses, the prerequisite relationship (which is either 0 or 1) between concepts might be obvious; however, the precise <test item, concept> relationships (which is a value ranging from 0 to 1) are difficult to determine. In the existing model, such relationships are determined based on the subjective opinions of individual teachers; therefore, different or even conflicting weightings might be given. Those inaccurate or controversial weights might significantly affect the accuracy of the learning diagnosis results. For example, in the same test sheet, if there are different <test item, concepts> weightings given by Expert *A* and Expert *B*, the learning diagnosis results provided by Expert *A* are PATH1: C_3 (0.35) → C_5 (0.6) and PATH2: C_5 (0.6), while those results given by Expert *B* are PATH1: C_3 (0.35) → C_5 (0.6) and PATH2: C_4 (0.4) → C_5 (0.6). That is, those inaccurate or controversial weights will affect the learning diagnosis results for a student. Therefore, it is important and challenging to find a method to check and integrate the weights given by multiple experts. This could be a means of decreasing inconsistencies in the weighting criteria.

3. A multi-expert approach for determining test item-concept relationships

Researchers have emphasized the importance of the collaboration of teachers in educating students (Blanton, Griffin, Winn, & Pugach, 1997; Nevin, Thousand, & Villa, 2007; Villa, Thousand, & Nevin, 2008). Nevin, Thousand, and Villa (2009) further argued that educators or teachers should interact with one another to share decision making, communication, and planning. It was found that teachers usually had their own experiences and would carefully propose and discuss their perspectives when collaborating with others; therefore, more effective outcomes could be attained. Moreover, researchers have shown that the idea of collaborative teaching plays important role not only in traditional teaching but also in computer-based learning; for example, the study presented by Hwang (2002) has demonstrated the benefits of collaboration of teachers in developing a computer-based tutoring environment.

In developing a diagnostic model for detecting the learning problems of students, it is difficult for a teacher to address all of the relationships between the test items and the concepts. From the educational perspective, the collaboration of teachers in determining such relationships could be an effective way to cope with this problem (Nevin et al., 2009). Consequently, it becomes an interesting and challenging issue to construct a set of rules to integrate the opinions of teachers. In this study, an expert system approach is employed to determine the weightings for each test item to the specified concepts by integrating the opinions of multiple experts.

Expert systems are defined as intelligent systems constructed by obtaining knowledge from human experts and coding it into a form that a computer may apply to similar problems (Hwang et al., 2009; Luger, 2005). Expert knowledge is a combination of a theoretical understanding of the problem and a collection of heuristic problem-solving rules that experience has shown to be effective in the domain (Hwang

et al., 2009; Luger, 2005). In the past decades, expert systems have been applied to not only many problem-solving applications, such as decision making, designing, planning, monitoring, diagnosing, and training activities (Chu & Hwang, 2008; Hwang et al., 2006; Liebowitz, 1997; Mahaman, Passam, Sideridis, & Yialouris, 2003; Zhou, Jiang, Yang, & Chen, 2002), but also to enhancing the learning process and improving the basic skills of students (Karake, 1990; Stankov, Rosić, Žitko, & Grubišić, 2008). Therefore, an expert system, the Knowledge Elicitation and Integration System for Determining the Weights of Concepts (KEISC), is developed accordingly. In KEISC, a multi-expert weight-presetting procedure (as shown in Fig. 2) is employed to elicit and integrate <test item, concept> relationships given by multiple experts. There are three phases of a multi-expert weight-presetting procedure, as follows: (1) the Elicitation of the weightings from the individual expert, (2) the Integration of the corresponding weightings, and (3) the Development testing and diagnostic system. In the following subsections, we shall introduce each phase of this procedure in detail.

Without loss of generality, in the following discussions, an integer value ranging from 0 to 5 (maximum-weighting) is used to denote "no relationship", "very weak", "weak", "average", "strong", and "very strong" relationships, respectively. In addition, the weighting for test item Q_i to concept C_j is represented as Weighting (Q_i, C_k), and the confidence degree for giving the weighting is represented as Certainty (Q_i, C_k). The values of Certainty (Q_i, C_k) could be either "*S*" or "*N*", where "*S*" represents "Sure" for giving the weighting, while "*N*" means "Not sure". Assume that n experts participate in the <test item, concept> determination process, the weighting given by Expert E_i for test item Q_j to concept C_k is represented as Weighting (E_i, Q_j, C_k), and the confidence degree for Expert E_i to give that value is represented as Certainty (E_i, Q_j, C_k). Furthermore, for convenience's sake, "*X*" is used to represent "no relationship". For example, if Expert *A* has determined that the weighting value for test item Q_1 to concept C_5 is 4 with high confidence, the relationship between the weighting and certainty values given by this expert can be represented as Weighting (E_A, Q_1, C_5) = 4 and Certainty (E_A, Q_1, C_5) = *S*. After integrating the Weighting (E_i, Q_j, C_k) values given by the experts, the corresponding TIRT (Q_j, C_k) value can be obtained via the formula TIRT (Q_j, C_k) = Weighting (Q_j, C_k)/maximum-weighting.

3.1. Eliciting <test item, concept> weightings from individual experts

This phase is invoked to elicit the <test item, concept> relationships of the test sheet from individual experts (experienced teachers). In this phase, each expert is asked to provide the weightings between the test item and the concept. Assume that the test sheet, which covers 4 concepts, contains 10 items. Expert A's opinions for determining the weighting values in this test sheet are shown in Table 2, where C_1, C_2, …, C_4 represent "Positive integer", "Addition of positive integer", "Subtraction of positive integer", and "Multiplication of positive integer", respectively.

3.2. Integrating corresponding weightings from multiple experts

In developing an expert system, one of the most difficult tasks is to gather domain knowledge from multiple experts. To cope with this problem, in this study, a set of rules in a knowledge base is defined to check and integrate the corresponding <test item, concept> relationships given by multiple experts. While interpreting these rules, we shall call the values that are less than 3 the "weak side", and those that are greater than 3 the "strong side". In addition, the value 3 is treated as being on both sides. There are four conditions for integrating the corresponding <test item, concept> relationships given by multiple experts as follows: (1) Integration rules for the same value with different degrees of confidence, (2) Integration rules for the values on the same side with different degrees of confidence, (3) Integration rules for the values with "*X*", and (4) Integration rules for the values on different sides. Hence, we shall introduce each condition of integrating corresponding <test item, concept> relationships in detail as follows:

- Integration rules for the same value with different confidence degrees
  Rule 1:
  If ∀i, (Weighting (E_i, Q_j, C_k) = v) and ∃p, ∋ Certainty (E_p, Q_j, C_k) = "*S*" or ∃q, ∋ Certainty (E_q, Q_j, C_k) = "*N*" and ((# of E_p) > (# of E_q)) where $1 \le i, p, q \le n$
  Then Weighting (Q_j, C_k) = v and Certainty (Q_j, C_k) = "*S*"

Rule 1 is used to handle the case that all of the experts have assigned the same value with different confidence degrees. In this case, it can be seen that these experts have an agreement on the weighting, and most of them show high confidence in making the decision. Therefore, v is adopted as the integrated weighting and the confidence degree is set to "*S*". Table 3 shows an illustrative example of integrating the opinions of five experts with this rule. By assuming that the test item is "Solve the Eq. (2)$X + 4 = 6$") and the experts are asked to provide the weighting value for describing the relationship between the test item and concept C_1 "Linear Equation with two variables". In this

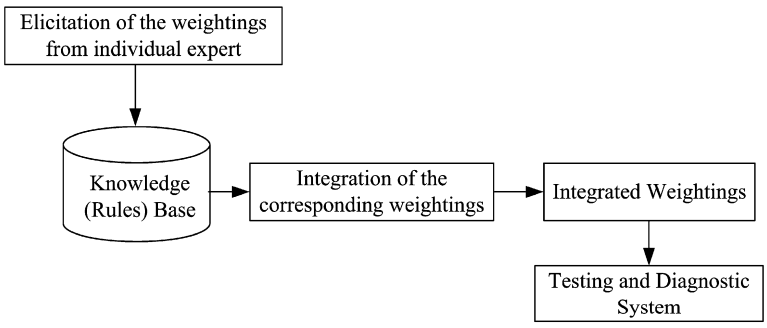

Fig. 2. Phases of the multi-expert weight-presetting procedure.

Table 2
Illustrative example of the <test item, concept> relationships provided by a single expert.

Test Q_i C_k	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
C_1	*X*,*S*	*X*,*S*	*X*,*S*	1,N	*X*,*S*	*X*,*S*	5,N	5,*S*	5,*S*	1,*S*
C_2	*X*,N	1,*N*	*X*,*S*	*X*,*S*	3,*S*	3,*N*	*X*,*S*	*X*,*S*	2,*S*	5,*S*
C_3	5,*S*	4,*S*	2,*S*	5,*S*	5,*S*	*X*,*S*	2,*S*	2,*S*	*X*,*S*	*X*,*S*
C_4	3,*S*	*X*,*N*	5,*S*	*X*,S	*X*,*S*	4,*S*	1,*S*	*X*,*N*	2,*S*	3,*S*

illustrative example, all of the experts have assigned "*X*" (no relationship) for relationship between the test item and the concept C_1, while three of them (E_1, E_2, and E_5) have high confidence and two of them (E_3 and E_4) have low confidence to represent the same relationship; therefore, the integrated weighting is "*X*" (no relationship) with confidence degree "*S*" (Sure). Another rule that is similar to this rule, but with most of the experts showing non-confidence in making their decision, is given as follows:

Rule 2:
If $\forall$i, (Weighting (E_i, Q_j, C_k) = v) and $\exists p, \ni$ Certainty (E_p, Q_j, C_k) = "*S*" or $\exists q, \ni$ Certainty (E_q, Q_j, C_k) = "*N*" and ((# of E_p) $\leq$ (# of E_q)) where $1 \leq i, p, q \leq n$
Then Weighting (Q_j, C_k) = v and Certainty (Q_j, C_k) = "*N*"

- Integration rules for the values on the same side with different confidence degrees

Rule 3:
If $\forall i$, (Weighting (E_i, Q_j, C_k) ≤ 3) and $\exists p, \ni$ Certainty (E_p, Q_j, C_k) = "*S*" or $\exists q, \ni$ Certainty (E_q, Q_j, C_k) = "*N*" and ((# of E_p) > (# of E_q)) where $1 \leq i, p, q \leq n$
Then Weighting (Q_j, C_k) = $\underset{\mathrm{i,\,i\,=\,p}}{\mathrm{MIN}}$(Weighting ($E_i$, Q_j, C_k)) and Certainty (Q_j, C_k) = "*S*"

Rule 3 is used to handle the case that all of the experts have assigned weak side values with different confidence degrees. In this case, it can be seen that the experts have an agreement on the weak side weighting, and most of them show high confidence in making the decision. Therefore, the smaller weighting with higher confidence is adopted as the integrated weighting, and the confidence degree is set to "*S*". Table 3 shows an illustrative example of integrating the opinions of five experts with this rule. By assuming that the test item is "Solve the equation $2X + 4 = 6$," and the experts are asked to provide the weighting value for describing the relationship between the test item and concept C_2 "Equation". In this illustrative example, all of the experts have assigned the weak side value, while three of them (E_1, E_3, and E_4) have confidence and two of them (E_2 and E_5) have low confidence to represent the same relationship; therefore, the integrated weighting is 1 by finding the minimum of values 1, 1 and 2, and the integrated confidence degree is "*S*" (Sure). Other rules that are similar to this rule are given as follows:

Table 3
Illustrative example of integrating the opinions of five experts in a set of rules.

Rule 1					
E_i	E_1	E_2	E_3	E_4	E_5
Weighting (E_i, Q_2, C_1)	*X*	*X*	*X*	*X*	*X*
Certainty (E_i, Q_2, C_1)	*S*	*S*	*N*	*N*	*S*
Weighting (Q_2, C_1) and Certainty (Q_2, C_1)	*X* and *S*				
Rule 3					
E_i	E_1	E_2	E_3	E_4	E_5
Weighting (E_i, Q_2, C_2)	1	2	1	2	3
Certainty (E_i, Q_2, C_2)	*S*	*N*	*S*	*S*	*N*
Weighting (Q_2, C_2) and Certainty (Q_2, C_2)	MIN(1, 1, 2) = 1 and *S*				
Rule 7					
E_i	E_1	E_2	E_3	E_4	E_5
Weighting (E_i, Q_2, C_3)	*X*	*X*	5	*X*	4
Certainty (E_i, Q_2, C_3)	*S*	*N*	*S*	*S*	*N*
Weighting (Q_2, C_3) and Certainty (Q_2, C_3)	Ask the experts to check and reconsider their weightings				
Rule 8a					
E_i	E_1	E_2	E_3	E_4	E_5
Weighting (E_i, Q_2, C_4)	*X*	3	1	2	*X*
Certainty (E_i, Q_2, C_4)	*N*	*S*	*S*	*S*	*N*
Weighting (Q_2, C_4) and Certainty (Q_2, C_4)	MIN(3, 1, 2) = 1 and *S*				
Rule 10a					
E_i	E_1	E_2	E_3	E_4	E_5
Weighting (E_i, Q_2, C_5)	3	*X*	1	2	*X*
Certainty (E_i, Q_2, C_5)	*N*	*N*	*N*	*N*	*N*
Weighting (Q_2, C_5) and Certainty (Q_2, C_5)	MIN (3, 1, 2) = 1 and *N*				
Rule 13					
E_i	E_1	E_2	E_3	E_4	E_5
Weighting (E_i, Q_2, C_6)	2	5	5	1	4
Certainty (E_i, Q_2, C_6)	*N*	*S*	*S*	*N*	*S*
Weighting (Q_2, C_6) and Certainty (Q_2, C_6)	MAX(5, 5, 4) = 5 and *S*				

Rule 4:
If $\forall i$, (Weighting (E_i, Q_j, C_k) ≤ 3) and $\exists p, \ni$ Certainty (E_p, Q_j, C_k) = "S" or $\exists q, \ni$ Certainty (E_q, Q_j, C_k) = "N" and ((# of E_p) $\leq$ (# of E_q)) where $1 \leq i, p, q \leq n$
Then Weighting (Q_j, C_k) = $\underset{i,\, i\, =\, p}{\text{MIN}}$ (Weighting (E_i, Q_j, C_k)) and Certainty (Q_j, C_k) = "N"

Rule 5:
If $\forall i$, (Weighting (E_i, Q_j, C_k) ≥ 3) and $\exists p, \ni$ Certainty (E_p, Q_j, C_k) = "S" or $\exists q, \ni$ Certainty (E_q, Q_j, C_k) = "N" and ((# of E_p) > (# of E_q)) where $1 \leq i, p, q \leq n$
Then Weighting (Q_j, C_k) = $\underset{i,\, i\, =\, p}{\text{MIN}}$ (Weighting (E_i, Q_j, C_k)) and Certainty (Q_j, C_k) = "S"

Rule 6:
If $\forall i$, (Weighting (E_i, Q_j, C_k) ≥ 3) and $\exists p, \ni$ Certainty (E_p, Q_j, C_k) = "S" or $\exists q, \ni$ Certainty (E_q, Q_j, C_k) = "N" and ((# of E_p) $\leq$ (# of E_q)) where $1 \leq$ i, $p, q \leq n$
Then Weighting (Q_j, C_k) = $\underset{i,\, i\, =\, p}{\text{MIN}}$ (Weighting (E_i, Q_j, C_k)) and Certainty (Q_j, C_k) = "N"

- Integration rules for the values with "X"

Rule 7:
If $\exists p, \ni$ (Weighting (E_p, Q_j, C_k) = "X" and Certainty (E_p, Q_j, C_k) = "S") and $\exists q, \ni$ ($1 \leq$ Weighting (E_q, Q_j, C_k) ≤ 5 and Certainty (E_q, Q_j, C_k) = "S"), where $1 \leq p, q \leq n$
Then Ask the experts to check and reconsider their ratings

Rule 7 is used to handle the case that some experts assign "X" (no relationship) with high confidence while others confidently assign an integer (having some degree of relationship) for the same <test item, concept> relationship. In this case, it can be seen that these experts have different opinions on determining the weighting; moreover, they show high confidence in making the decision. Therefore, the system will return to ask the experts to check and reconsider their ratings. Table 3 shows an illustrative example of this rule; that is, two experts (E_1 and E_4) have assigned "X" with high confidence, and one expert (E_3) has assigned an integer to represent the same relationship with high confidence; that is, a conflicting opinion exists. In this case, the system will ask them to check and reconsider their ratings.

Rule 8a:
If $\exists p, \ni$ (Weighting (E_p, Q_j, C_k) = "X" and Certainty (E_p, Q_j, C_k) = "N") and $\exists q, \ni$ ($1 \leq$ Weighting (E_q, Q_j, C_k) ≤ 3 and Certainty (E_q, Q_j, C_k) = "S") where $1 \leq p, q \leq n$
Then Weighting (Q_j, C_k) = $\underset{i,\, i\, =\, p}{\text{MIN}}$ (Weighting (E_i, Q_j, C_k)) and Certainty (Q_j, C_k) = "S"

Rule 8a is used to handle the case that some experts assigned "X" (no relationship) with less confidence while some confidently assigned an integer on the weak side value (having some degree of relationship) for the same <test item, concept> relationship. In this case, these experts show different opinions in determining the weighting with different confidence degrees. Therefore, the smaller weighting with higher confidence is adopted as the integrated weighting, and the confidence degree is set to "S". An illustrative example of this rule is given in Table 3, in which two experts (E_1 and E_5) have not confidently assigned "X", and three experts (E_2, E_3, and E_4) have assigned an integer on the weak side value to represent the same relationship with high confidence; therefore, the integrated rating is equal to 1 by finding the minimum of values 3, 1 and 2, and the integrated confidence degree is "S" (Sure). Another rule that is similar to this rule is given as follows:

Rule 8b:
If $\exists p, \ni$ (Weighting (E_p, Q_j, C_k) = "X" and Certainty (E_p, Q_j, C_k) = "N") and $\exists q, \ni$ ($3 \leq$ Weighting (E_q, Q_j, C_k) ≤ 5 and Certainty (E_q, Q_j, C_k) = "S") where $1 \leq p, q \leq n$
Then Weighting (Q_j, C_k) = $\underset{i,\, i\, =\, p}{\text{MIN}}$ (Weighting (E_i, Q_j, C_k)) and Certainty (Q_j, C_k) = "S"

Rule 9a:
If $\exists p, \ni$ (Weighting (E_p, Q_j, C_k) = "X" and Certainty (E_p, Q_j, C_k) = "S") and $\exists q, \ni$ ($1 \leq$ Weighting (E_q, Q_j, C_k) ≤ 3 and Certainty (E_q, Q_j, C_k) = "N") where $1 \leq p, q \leq n$
Then Weighting (Q_j, C_k) = "X" and Certainty (Q_j, C_k) = "S"

Rule 9a is used to handle the case that some experts assign "X" (no relationship) with high confidence while some assign weak side values with less confidence. In this case, "X" is adopted as the integrated weighting, and the confidence degree is set to "S". Another rule that is similar to this rule is given as follows:

Rule 9b:
If $\exists p, \ni$ (Weighting (E_p, Q_j, C_k) = "X" and Certainty (E_p, Q_j, C_k) = "S") and $\exists q, \ni$ ($3 \leq$ Weighting (E_q, Q_j, C_k) ≤ 5 and Certainty (E_q, Q_j, C_k) = "N") and where $1 \leq p, q \leq n$
Then Weighting (Q_j, C_k) = "X" and Certainty (Q_j, C_k) = "N"

- Integration rules for the values on different sides

Rule 10a:
If $\exists p, \ni$ (Weighting (E_p, Q_j, C_k) = "X" and Certainty (E_p, Q_j, C_k) = "N") and $\exists q, \ni$ ($1 \leq$ Weighting (E_q, Q_j, C_k) ≤ 3 and Certainty (E_q, Q_j, C_k) = "N") and $\sim\exists r, \ni$ Certainty (E_r, Q_j, C_k) = "S" where $1 \leq p, q, r \leq n$
Then Weighting (Q_j, C_k) = $\underset{i,\, i\, =\, p}{\text{MIN}}$ (Weighting (E_i, Q_j, C_k)) and Certainty (Q_j, C_k) = "N"

Rule 10a is used to handle the case that some experts assign "*X*" (no relationship) with less confidence, while others assign the weak side values with less confidence. In this case, it can be seen that these experts have different opinions on determining the weighting; moreover, they have no confidence in making the decision. Therefore, the smaller weighting is adopted as the integrated weighting, and the confidence degree is set to "*N*". Table 3 shows an illustrative example of this rule. Assume that two experts (E_2 and E_5) have assigned "*X*" with non-confidence, and three experts (E_1, E_3, and E_4) have assigned an integer on the weak side value to represent the same relationship with non-confidence; therefore, the integrated weighting is equal to 1 by finding the minimum of values 3, 1 and 2, and the integrated confidence degree is "*N*" (Not Sure). Another rule that is similar to this rule is given as follows:

Rule 10b:

If $\exists p, \ni$ (Weighting (E_p, Q_j, C_k) = "*X*" and Certainty (E_p, Q_j, C_k) = "*N*") and $\exists q, \ni$ ($3 \leq$ Weighting (E_q, Q_j, C_k) ≤ 5 and Certainty (E_q, Q_j, C_k) = "*N*") and ~$\exists r, \ni$ Certainty (E_r, Q_j, C_k) = "*S*" where $1 \leq p, q, r \leq n$

Then Weighting (Q_j, C_k) = $\underset{i,\, i\, =\, p}{\text{MIN}}$(Weighting ($E_i$, Q_j, C_k)) and Certainty (Q_j, C_k) = "*N*"

Rule 11:

If $\exists p, \ni$ Weighting (E_p, Q_j, C_k) < 3 and Certainty (E_p, Q_j, C_k) = "*S*" and $\exists q, \ni$ Weighting (E_q, Q_j, C_k) > 3 and Certainty (E_q, Q_j, C_k) = "*S*", where $1 \leq p, q \leq n$

Then Ask the experts to check and reconsider their ratings

Rule 11 handles the case that some experts assign an integer that is less than 3 with high confidence, while some confidently assign an integer that is greater than 3 to express the same <test item, concept> relationship. In this case, it can be seen that these experts have different opinions on determining the weighting; moreover, they show high confidence in making the decision. Therefore, the system will return to ask the experts to check and reconsider their ratings. Another rule that is similar to this rule is given as follows:

Rule 12:

If $\exists p, \ni$ Weighting (E_p, Q_j, C_k) < 3 and Certainty (E_p, Q_j, C_k) = "*N*" and $\exists q, \ni$ Weighting (E_q, Q_j, C_k) > 3) and Certainty (E_q, Q_j, C_k) = "*N*" where $1 \leq p, q \leq n$

Then Ask the experts to check and reconsider their ratings

Rule 13:

If ~$\exists p, \ni$ Weighting (E_p, Q_j, C_k) = "*X*" and $\exists q, \ni$ Weighting (E_q, Q_j, C_k) < 3 and Certainty (E_q, Q_j, C_k) = "*N*" and $\exists r, \ni$ Weighting (E_r, Q_j, C_k) > 3 and Certainty (E_r, Q_j, C_k) = "*S*" where $1 \leq p, q, r \leq n$

Then Weighting (Q_j, C_k) = $\underset{i,\, i\, =\, p}{\text{MIN}}$(Weighting ($E_i$, Q_j, C_k) and Certainty (Q_j, C_k) = "*S*"

Rule 13 handles the case that some experts assign an integer that is less than 3 with non-confidence, while some confidently assign an integer that is greater than 3 to express the same <test item, concept> relationship with high confidence. In this case, it can be seen that these experts have different opinions on determining the weighting; moreover, they show different confidence in making the decision. Therefore, the larger weighting with higher confidence is adopted as the integrated weighting, and the confidence degree is set to "*S*". Table 3 shows an illustrative example of this rule, where two experts (E_1 and E_4) have assigned weights that are less than 3 with non-confidence, and three experts (E_2, E_3, and E_5) have assigned values that are greater than 3 to represent the same relationship with high confidence; therefore, the integrated weighting is equal to 5 by finding the maximum of values 5, 5 and 4, and the integrated confidence degree is "*S*" (Sure). Another rule that is similar to this rule is given as follows:

Rule 14:

If ~$\exists p, \ni$ Weighting (E_p, Q_j, C_k) = "*X*" and $\exists q, \ni$ Weighting (E_q, Q_j, C_k) ≤ 3 and Certainty (E_q, Q_j, C_k) = "*S*" and $\exists r, \ni$ Weighting (E_r, Q_j, C_k) ≥ 3 and Certainty (E_r, Q_j, C_k) = "*N*" where $1 \leq p, q, r \leq n$

Then Weighting (Q_j, C_k) = $\underset{i,\, i\, =\, p}{\text{MIN}}$(Weighting ($E_i$, Q_j, C_k)) and Certainty (Q_j, C_k) = "*S*"

The integration <test item, concept> weights from the multiple expert phase is repeatedly conducted until no further checking and considering weighting information is needed.

3.3. Developing testing and diagnostic system

The integrated <test item, concept> weightings from multiple experts are used to diagnose student learning problems by developing a testing and diagnostic system. A diagnostic learning method is used to diagnose learning problems for students according to their own test answers. The personalized learning guidance was then provided to each student.

4. System development

Based on our novel approach, a testing and diagnosis system has been implemented using the *C*#.net programming language. The structure of the system is depicted in Fig. 3.

In the expert system, KEISC, a weighting value interface is provided to allow teachers to elicit the <item, concept> relationships (Fig. 4). Without loss of generality, we have selected one of the most popular expert system shells, the C language integrated production system (CLIPS), as the target rule format.

After all of the experts determine the <test item, concept> weightings for some specific application domains, KEISC integrates the weightings elicited from individual experts. If there are some conflicts occurring after integration, KEISC will require all experts to check and

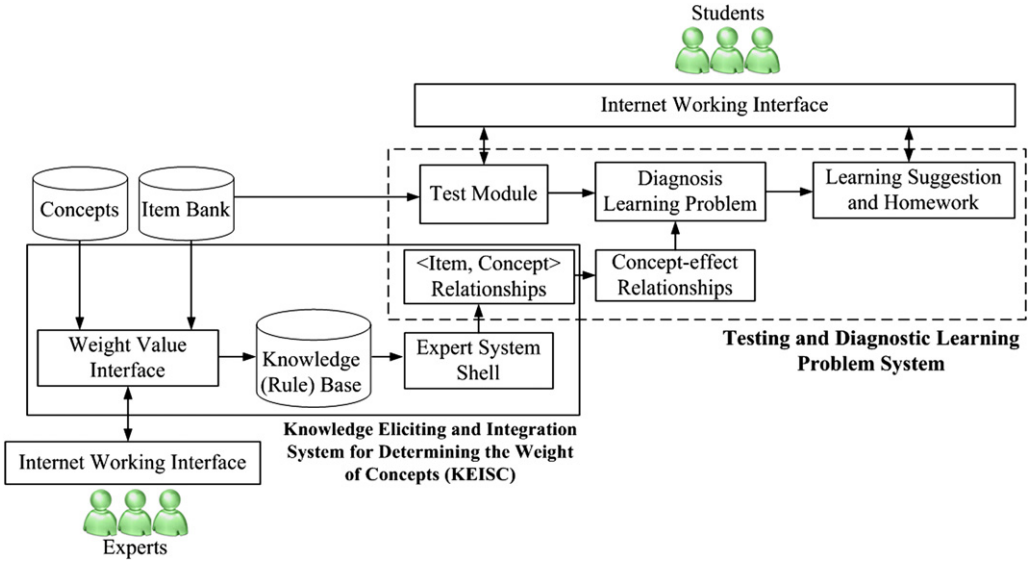

Fig. 3. Structure of the testing and diagnostic system.

reconsider their <test item, concept> weightings and certainty values as shown in Fig. 5. KEISC is repeatedly conducted until no further checking and considering of the <test item, concept> weighting information is needed. The integrated <test item, concept> weightings from KEISC have been used as one of the inputs into the Testing and Diagnostic Learning Problem (TDLP).

In TDLP, a learning diagnosis method is then used to detect student learning problems. The students are requested to log on to their computers. The testing system will generate a test sheet for the students according to the concepts specified by the teacher. After the

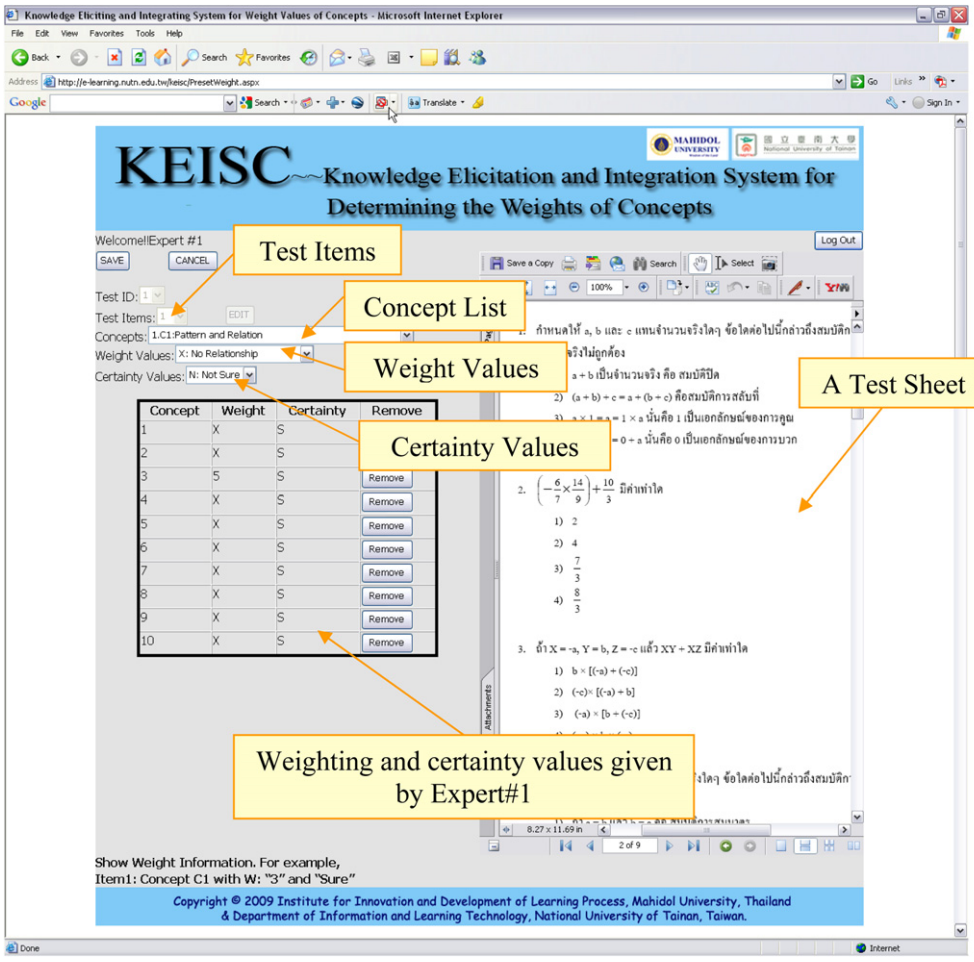

Fig. 4. Illustrative example of the weighting value interface.

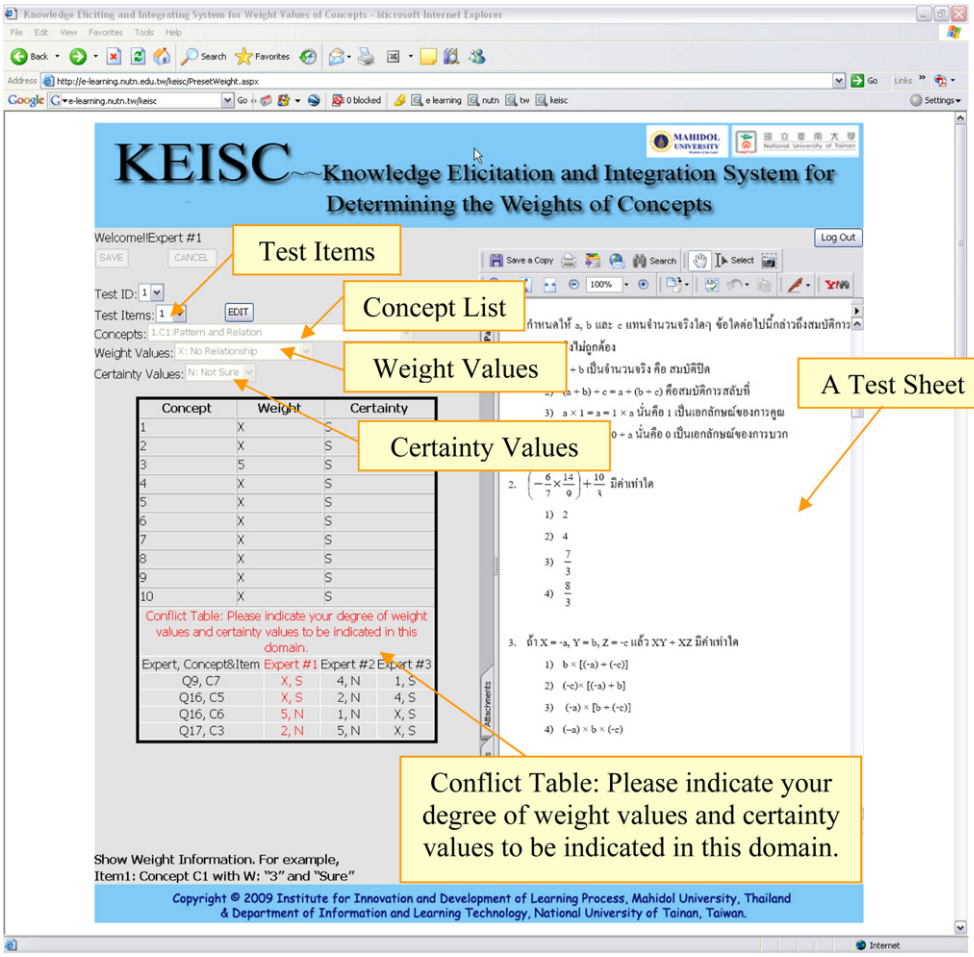

Fig. 5. KEISC indicates that reconsidering weighting is needed.

students submit their answers, the system will analyze the answers and provide personalized learning guidance based on the concept-effect relationships. Fig. 6 shows an illustrative example of learning guidance as well as a summarized test report for a student. Moreover, those data are used to arrange homework for each student to improve their learning achievements.

5. Evaluation and analysis

To find out if the innovative approach would be helpful in achieving learning achievement, system evaluation and analysis need to be conducted.

5.1. Experiment design

To evaluate the performance of this innovative approach, an experiment has been conducted. Three experienced teachers, each with 15 years experience teaching mathematics courses, served as the domain experts. A total of 113 students at a junior high school in Thailand, including fifty-four females and fifty-nine males with an average age of 15, participated in the computer-based mathematics course on the topic "System of Linear Equations". These students were divided into four groups (i.e., Control Group 1, Control Group 2, Control Group 3, and Experimental Group 1) to compare the performance of the innovative approach and the original CER model.

(1) Control group CG_1: In this group, 31 students received learning suggestions from the CER model with the <test item, concept> weightings given by domain expert#1.
(2) Control group CG_2: In this group, 31 students received learning suggestions from the CER model with the <test item, concept> weightings given by domain expert#2.
(3) Control group CG_3: In this group, 25 students received learning suggestions from the CER model with the <test item, concept> weightings given by domain expert#3.
(4) Experimental group E_1: In this group, 26 students received learning guidance from the enhanced learning diagnosis model that integrated <test item, concept> weightings from multiple experts.

All of the students took a pre-test, three self-assessments, and a post-test. Firstly, the students took a pre-test to evaluate whether they had equivalent mathematical background knowledge prior to taking the course. According to the pre-test answers, the students in the

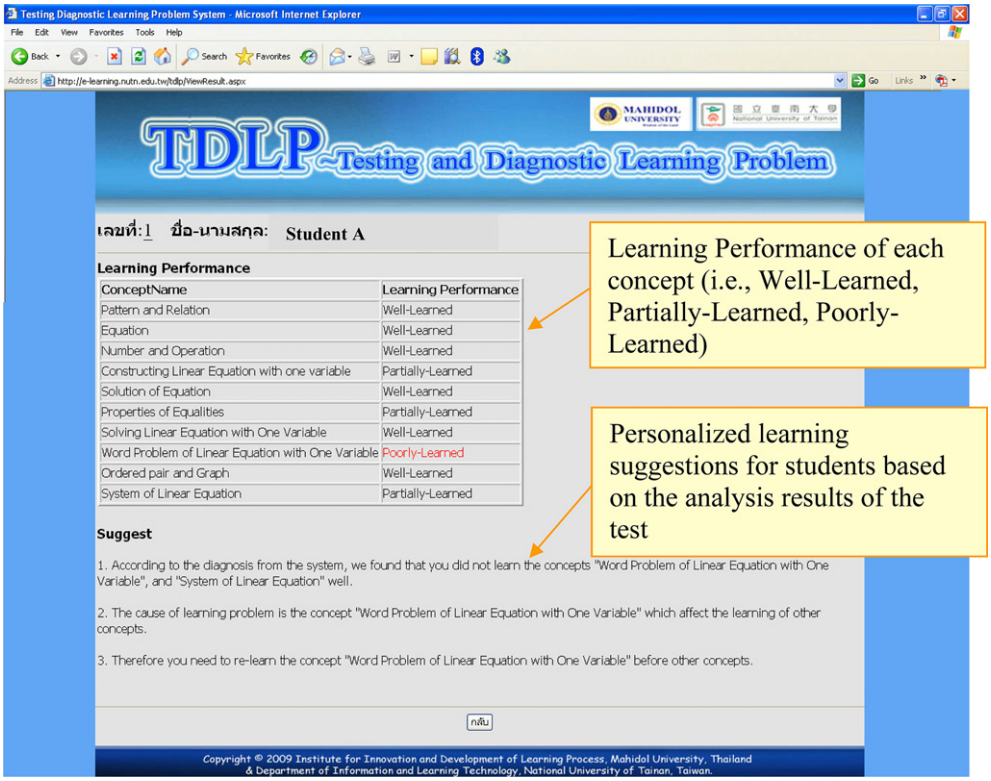

Fig. 6. Illustrative example of a personalized learning suggestion.

control groups (i.e., CG_1, CG_2 and CG_3) received the learning suggestions from the original CER model and relevant homework, while those in the experimental group (i.e., E_1) received the learning guidance generated by applying our innovative approach, and relevant homework. The experiment was conducted in August, 2009. The entire learning activity lasted four weeks, during which the students took three tests. The test results were analyzed by applying the learning diagnostic system for generating learning guidance and assigning corresponding homework to individual students. After finishing the learning activity, all of the students took a post-test. In the following, the statistical tests are employed to test the difference statistically among all participating groups.

5.2. Analysis of pre-test

The pre-test aimed to ensure that all groups of students had the equivalent mathematics knowledge required for taking the System of Linear Equations course. The test sheet of the pre-test consisted of 30 multiple-choice test items concerning the basic knowledge related to the study of the system of linear equations. Each item was scored one point. The KR-20 reliability of the pre-test was 0.86. The item difficulty index was around 0.3–0.7 with an average difficulty degree of 0.49, which was close to the optimum value 0.5. The item discrimination index of most items was greater than 0.3, implying that the items did discriminate usefully (Doran, 1980).

Prior to the ANOVA test, the Levene's test of homogeneity of variances was applied to examine whether variances across samples were equal. The result of this test was not significant ($p = .342 > .05$), which suggests that the difference between the variances for all groups was not significant. Therefore, ANOVA was performed. The results of analyzing the pre-test showed that the average pre-test scores of the students in the four participating groups were equal ($F(3, 109) = .925$, $p = .431 > .05$). These results indicate that there was no statistically significant difference between the mean scores for the groups. Consequently, we concluded that the four groups of students had an equivalent level of knowledge prior to taking the System of Linear Equations course.

5.3. Analysis of learning achievement

The post-test was conducted to compare the linear equations knowledge between the four groups of students after receiving the learning suggestions. In the post-test, the KR-20 reliability was 0.84. The item difficulty index was around 0.3 to 0.7, with an average difficulty degree of 0.48, which was close to the optimum value of 0.5. The item discrimination index of most items was greater than 0.3, implying that the items did discriminate usefully (Doran, 1980).

First, the Levene's test for equality of variances was not significant ($p = .343 > .05$), which indicates that the variances for all groups were assumed to be equal. A one-way ANOVA was conducted, revealing that there was significant difference in the post-test scores of the four groups of participating students, $F(3, 109) = 42.783$, $p = .000 < .05$; at least one contrast was significantly different. The Scheffe test was used to make post hoc comparisons to demonstrate where we can find out the statistically significant difference between the four groups of participating students. Table 4 shows the significance level for these multiple comparisons, taking into account that the significance level for the mean difference is $p < .05$, which has been marked with an asterisk "*".

Table 4
Means, standard deviations, and One-Way Analyses of Variance (ANOVA) summary table for the post-test results of the four groups.

	Group	Number of students	Average score	Standard deviation	$F(3, 109)$	Post hoc test (Scheffe)
(a)	E_1	26	24.96	3.594	42.783	(a) > (b)*
(b)	CG_1	31	16.39	3.774		(a) > (c)*
(c)	CG_2	31	16.26	2.792		(a) > (d)*
(d)	CG_3	25	17.72	2.880		

*$p < .05$.
(a) Experimental group E_1: the enhanced learning diagnosis model's suggestions given by multiple experts (b) Control group CG_1: CER model's suggestions given by domain expert#1; (c) Control group CG_2: CER model's suggestions given by domain expert#2; (d) Control group CG_3: CER model's suggestions given by domain expert#3.

From this table, it is clear that there are significant differences between Control group CG_1 and Experimental group E_1, between Control group CG_2 and Experimental group E_1, and Control group CG_3 and Experimental group E_1. Consequently, these results indicate that the students in experimental group (E_1) achieved significantly better performance than those in the three control groups after implementing the innovative approach, at the confidence interval of 95%.

Although our innovative approach is able to achieve significantly better achievement than the original CER model, for the same concept, the learning suggestions provided by the single-experts differed from those given by the multiple experts. Therefore, another analysis was made to compare the learning improvement of the students in the four groups (i.e., CG_1, CG_2, CG_3, and E_1). Table 5 shows the *t*-test results of the learning improvement for these groups by analyzing the pre- and post-test results. The results of learning improvement (post-test vs. pre-test) were not significant among the control group students (i.e., CG1: $t = 0.090$, $p > .05$; CG2: $t = 1.374$, $p > .05$; CG3: $t = -0.799$, $p > .05$). It is clear that the students in the three control groups (i.e., CG_1, CG_2, and CG_3) did not improve their learning performance after using the learning suggestions from the original CER model. In contrast, in the experimental group (i.e., E_1), the analysis of the pre- and post-test results suggests that there was significant difference between the pre- and post-test mean scores ($t = -6.376$, $p < .05$). Obviously, the students in the experimental group improved their learning achievement significantly after using the diagnosis learning results from our innovative approach.

From the analysis of the experimental data, two observations can be derived:

(1) From the post-test results, the enhanced learning diagnosis model's suggestions given by multiple experts can improve the learning achievement of students in comparison with the CER model's suggestions given by a single expert.
(2) From a comparison of the pre-test and post-test results, it can be seen that only those students who received the learning suggestions from our innovative approach significantly improved their learning achievement.

To sum up, we conclude that our innovative approach is particularly helpful to students in enhancing their learning achievement. There is an important reason for the improvement in the learning achievement of the students after implementing our innovative approach. That is, if there are some conflicts occurring after integrating <test item, concept> weightings when using the multi-expert systems, the system requires all experts to check and reconsider their weightings and certainty values. This implies that inconsistent weightings are rejected and more consistent weightings are determined for the system. This also minimizes inconsistent weighting by each expert. Therefore, there are high quality weightings associated with the multi-expert system resulting in more accurate learning suggestions being provided to the students in the experimental group than those in the three control groups. For this reason, the students who received the learning guidance given by our innovative approach improved their learning achievement in comparison with those who received the learning suggestions from the original CER model.

6. Conclusions and discussions

This study proposes a multi-expert approach for eliciting and integrating the weightings of test item-concept relationships from multiple experts. A testing and diagnostic system has been implemented and an experiment has been conducted to evaluate the performance of the proposed approach. In the following subsections, detailed discussions are given to explain how this research contributes to the field of computers and education and how this study innovates and extends the current literature. Moreover, the matters to be noted, the step-by-step procedure to be followed and the limitations of using the approach are given to instruct researchers and practitioners who want to apply the proposed approach.

Table 5
t-Test results of learning improvement for the four groups of students.

Group	Tests	*N*	Mean	S.D.	*T*
E_1	Pre-test	26	18.31	4.155	−6.376*
	Post-test	26	24.96	3.594	
CG_1	Pre-test	31	16.48	4.523	0.090
	Post-test	31	16.39	3.774	
CG_2	Pre-test	31	17.29	4.368	1.374
	Post-test	31	16.26	2.792	
CG_3	Pre-test	25	17.12	3.270	−0.799
	Post-test	25	17.72	2.880	

*$p < .05$.

6.1. Contribution of the proposed approach to educational technology areas

To improve the learning achievement of students, it is important to diagnose the learning problems and provide learning suggestions for individual students. In the past decade, several studies have been conducted for diagnosing student learning problems and providing appropriate learning guidance for individual students. Among the existing models, the CER model (Hwang, 2003a) has been widely adopted by researchers for detecting the learning barriers of students and improving their learning achievements in several fields, including Natural Science, Mathematics, Physics, Electronic Engineering, and Health (Chu, Hwang, Tseng, & Hwang, 2006; Hwang et al., 2007; Jong et al., 2004; Tseng et al., 2007).

Nevertheless, although the CER model seems to be effective in several applications, past experiences of using this model have also revealed a drawback (Hwang, Tseng et al., 2008; Lee et al., 2009); that is, individual teachers might provide some imprecise <test item, concept> relationships owing to ignorance or subjectivity; moreover, researchers have indicated that domain experts with different experiences could have different expertise or understanding of each portion of the knowledge; therefore, the cooperation of several experts (experienced teachers) has been suggested (Chu & Hwang, 2008; Huang & Shimizu, 2006; Hwang et al., 2006; Léger & Naud, 2009; Medsker et al., 1995; Mittal & Dym, 1985). Yang and Chan (2008) further indicated that a multi-expert approach is an effective way to recognize and reject incorrect solutions and suggestions to a higher degree than a single-expert approach.

However, so far little research has been devoted to investigating this important and challenging issue, that is, the cooperation of multiple experts for obtaining more accurate ratings to describe the relevance between the test items and the concepts to be assessed.

The major contribution of this study is to propose a multi-expert approach to elicit and integrate <test item, concept> relationships from multiple experts, such that more accurate learning guidance or suggestions can be provided to individual students. The performance of this innovative approach has been compared with that of the existing model by conducting an experiment on a Mathematics course in a junior high school. From the cooperation of three experienced teachers and the test results of 113 junior high school students, it is found that the proposed approach is able to provide more effective learning guidance to the students; that is, the students who follow the proposed approach are able to achieve significantly better performance than those who received the suggestions given by the previously proposed model. Therefore, we conclude that the proposed approach can cope with the key barrier of diagnosing student learning problems. Such a contribution not only plays an important role in enhancing the learning performance of students, but also provides valuable references for researchers who are devoted to the study of models of learning diagnosis (Hwang et al., 2003; Jong et al., 2004) or the development of testing and diagnostic systems (Chen & Bai, 2009; Chu et al., 2006; Drew, Thorpe, & Bannisher, 2000; Hwang, 2007; Lee et al., 2009).

6.2. Guidelines for researchers and practitioners to apply the proposed approach

The proposed approach can be applied to most computer-based courses, such as Natural Science, Physics, Chemistry and Mathematics, which contain explicit <test item, concept> relationships and concept-effect relationships. When applying this approach, it is suggested that three to seven experienced teachers be invited to participate in the determination of those <test item, concept> relationships and the concept-effect relationships (Chu & Hwang, 2008; Huang & Shimizu, 2006; Léger & Naud, 2009).

In developing learning diagnosis systems with this model, several functions need to be taken into consideration:

(1) A <test item, concept> management unit that can assist the researchers to elicit and manage <test item, concept> relationships from the teachers and integrate their opinions. It is important to know that most teachers might have difficulty in realizing the physical meanings of the ratings; therefore, it would be better to show the relationships with a description like “highly relevant” or “irrelevant” instead of a rating value like “5” or “2” when designing the user interface.
(2) A graphical concept-effect relationship function that allows teachers to develop the concept-effect relationships via discussing and drawing the graph cooperatively.
(3) A teacher interface that allows individual teachers to log in, define thresholds for analyzing the learning problems of the students in their classes, and initiate the learning diagnosis procedure.
(4) A testing unit that allows students to log in to receive tests online.
(5) A test result importing unit that allows the teachers to upload the test results for analysis if the tests are conducted via the traditional paper-and-pencil approach.
(6) A learning diagnosis unit that analyzes the students learning problems based on the integrated <test item, concept> relationships and the CER model. The learning guidance generated by the system can be browsed online through the teacher accounts or the student accounts. The students are only allowed to browse their own learning diagnosis results, while the teachers can browse the information for all of the students in their classes. The diagnosis results can also be downloaded and printed out.

In addition, to introduce the proposed approach into the classroom, a step-by-step lead-in procedure is given as follows:

Step 1: Explain the meanings of concepts, <test item, concept> relationship and concept-effect relationship to the teachers by showing some examples. This step will take 30–60 min.
Step 2: Show the learning diagnosis system to the teachers. Usually it takes 20–30 min to demonstrate the functions of the system.
Step 3:Guide the teachers to determine the concepts (or learning objectives) involved in the target subject unit (the scope for conducting the learning diagnosis experiment). Usually the text books have a guideline that shows all of the concepts or learning objectives. This step can be carried out online or offline, depending on the practical needs.
Step 4: Ask the teachers to provide a set of test items for the target subject unit. Remove the redundant or similar test items upon the permission of the teachers. This step can also be carried out online or offline, depending on the practical needs.
Step 5: Guide individual teachers to determine the <test item, concept> relationships. This step is usually done online.
Step 6: Guide the teachers to determine the concept-effect relationships. This step is usually done online using the graphical interface or a video conference.

Step 7: Conduct the online tests or paper-and-pencil tests. For the students who receive paper-and-pencil tests, the test results need to be input and uploaded via the test result importing unit.
Step 8: Analyze the test results and generate the learning guidance for each student. In this step, some supplemental materials can be provided to the students online if there are some available learning systems. Alternatively, the teachers can ask the students to re-study the relevant part in the text book or provide additional homework to individual students based on the learning guidance.

For researchers who would like to compare the students' learning achievements before and after receiving the learning guidance, an additional post-test is needed. If the objective of the research is to investigate the effectiveness of the treatment given in Step 8, control groups might be needed to compare the performance of the students who receive different treatments in this step.

Furthermore, for those researchers or teachers who have difficulty in implementing their own learning diagnosis systems, the authors are willing to provide accounts of the system, program codes, or technical assistance upon request.

6.3. Limitation of the approach

Although our new approach reveals good performance, it has some limitations in its practical application and now is on the verge of being improved and optimized. For example, the analysis results could be inaccurate owing to correct answers due to "lucky guesses", or incorrect answers due to "carelessness". To avoid these situations, it is suggested that the teachers design more than one test item for each concept, depending on the available testing time. In addition, a two-tiered test, which employs two-level multiple-choice questions to diagnose students' alternative conceptions, could be an alternative to cope with this problem (Tsai & Chou, 2002).

Moreover, it would be interesting to know whether the same approach would work, and how well, for other kinds of courses, such as Natural Science, Science and Engineering courses. Consequently, further investigations have been planned to apply this novel approach to online tutoring for different courses.

Acknowledgments

This study is supported in part by the Institute for the Promotion of Teaching Science and Technology (IPST), the Thailand Research Fund, the Commission on Higher Education, and the National Science Council of the Republic of China under contract numbers NSC 98-2511-S-024-007-MY3 and NSC 98-2631-S-024-001.

References

Bai, S. M., & Chen, S. M. (2008). Automatically constructing concept maps based on fuzzy rules for adapting learning systems. *Expert Systems with Applications, 35*(1), 41–49.
Blanton, L., Griffin, C., Winn, J., & Pugach, M. (1997). *Teacher education in transition: Collaborative programs to prepare general and special educators*. Denver, CO: Love.
Casamayor, A., Amandi, A., & Campo, M. (2009). Intelligent assistance for teachers in collaborative e-learning environments. *Computers & Education, 53*(4), 1147–1154.
Chen, S. M., & Bai, S. M. (2009). Learning barriers diagnosis based on fuzzy rules for adaptive learning systems. *Expert Systems with Applications, 36*(8), 11211–11220.
Cheng, S. Y., Lin, C. S., Chen, H. H., & Heh, J. S. (2005). Learning and diagnosis of individual and class conceptual perspectives: an intelligent systems approach using clustering techniques. *Computers & Education, 44*(3), 257–283.
Chiou, C. K., Hwang, G. J., & Tseng, J. C. R. (2009). An auto-scoring mechanism for evaluating problem-solving ability in a web-based learning environment. *Computers & Education, 53*(2), 261–272.
Chu, H. C., & Hwang, G. J. (2008). A Delphi-based approach to developing expert systems with the cooperation of multiple experts. *Expert Systems with Applications, 34*(4), 2826–2840.
Chu, H. C., Hwang, G. J., Tseng, J. C. R., & Hwang, G. H. (2006). A computerized approach to diagnosing student learning problems in health education. *Asian Journal of Health and Information Sciences, 1*(1), 43–60.
Doran, R. (1980). *Basic measurement and evaluation of science instruction*. Washington D.C.: National Science Teachers Association.
Drew, S., Thorpe, L., & Bannisher, P. (2000). Key skills computerized assessments: guiding principles. *Assessment and Evaluation in Higher Education, 27*(2), 175–186.
Gerber, M., Grund, S., & Grote, G. (2008). Distributed collaboration activities in a blended learning scenario and the effects on learning performance. *Journal of Computer Assisted Learning, 24*(3), 232–244.
Gronlund, N. E. (2003). *Assessment of student achievement*. USA: Pearson Education.
Hooper, S. (1992). Cooperative learning and computer-based instruction. *Educational Technology Research & Development, 40*(3), 21–38.
Huang, C. J., Liu, M. C., Chu, S. S., & Cheng, C. L. (2007). An intelligent learning diagnosis system for web-based thematic learning platform. *Computers & Education, 48*(4), 658–679.
Huang, L. L., & Shimizu, A. (2006). A multi-expert approach for robust face detection. *Pattern Recognition, 39*(9), 1695–1730.
Huang, Y. M., Lin, Y. T., & Cheng, S. C. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers & Education, 52*(1), 53–67.
Hwang, G. H., Chen, J. M., Hwang, G. J., & Chu, H. C. (2006). A time scale-oriented approach for building medical expert systems. *Expert Systems with Applications, 31*(2), 299–308.
Hwang, G. J. (2002). On the development of a cooperative tutoring environment on computer networks. *IEEE Transaction on Systems, Man, and Cybernetics, 32*(3), 272–278.
Hwang, G. J. (2003a). A conceptual map model for developing intelligent tutoring systems. *Computers & Education, 40*(3), 217–235.
Hwang, G. J. (2003b). A test-sheet-generating algorithm for multiple assessment requirements. *IEEE Transactions on Education, 46*(3), 329–337.
Hwang, G. J. (2007). Gray forecast approach for developing distance learning and diagnostic systems. *IEEE Transaction on Systems, Man, and Cybernetics C, 37*(1), 98–108.
Hwang, G. J., Cheng, H., Chu, C. H. C., Tseng, J. C. R., & Hwang, G. H. (2007). Development of a web-based system for diagnosing student learning problems on English tenses. *Journal of Distance Education Technologies, 5*(4), 80–98.
Hwang, G. J., Chu, H. C., Yin, P. Y., & Lin, J. Y. (2008). An innovative parallel test-sheet composition approach to meet multiple assessment criteria for national tests. *Computers & Education, 51*(3), 1058–1072.
Hwang, G. J., Hsiao, C. L., & Tseng, J. C. R. (2003). A computer-assisted approach to diagnosing student learning problems in science courses. *Journal of Information Science and Engineering, 19*(2), 229–248.
Hwang, G. J., Tsai, P. S., Tsai, C. C., & Tseng, J. C. R. (2008). A novel approach for assisting teachers in analyzing student web-searching behaviors. *Computers & Education, 51*(2), 926–938.
Hwang, G. J., Tseng, J. C. R., & Hwang, G. H. (2008). Diagnosing student learning problems based on historical assessment records. *Innovations in Education and Teaching International, 45*(1), 77–89.
Hwang, G. J., Yang, T. C., Tsai, C. C., & Yang, S. J. H. (2009). A context-aware ubiquitous learning environment for conducting complex experimental procedures. *Computers & Education, 53*(2), 402–413.
Jong, B. S., Chan, T. Y., & Wu, Y. L. (2007). Learning log explorer in e-learning diagnosis. *IEEE Transactions on Education, 50*(3), 216–228.
Jong, B. S., Lin, T. W., Wu, Y. L., & Chan, T. Y. (2004). Diagnostic and remedial learning strategy based on conceptual graphs. *Journal of Computer Assisted Learning, 20*(5), 377–386.
Karake, Z. A. (1990). Enhancing the learning process with expert systems. *Computers & Education, 14*(6), 495–530.

Lee, C. H., Lee, G. G., & Leu, Y. (2009). Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning. *Expert Systems with Applications, 36*(2), 1675–1684.
Léger, B., & Naud, O. (2009). Experimenting statecharts for multiple experts knowledge elicitation in agriculture. *Expert Systems with Applications, 36*(8), 11296–11303.
Liebowitz, J. (1997). Knowledge-based/expert systems technology in life support systems. *International Journal of Systems & Cybernetics, 26*(5), 555–573.
Luger, G. F. (2005). *Artificial intelligence: Structures and strategies for complex problem solving* (5th ed.). London: Addison-Wesley.
Mahaman, B. D., Passam, H. C., Sideridis, A. B., & Yialouris, C. P. (2003). DIARES-IPM: a diagnostic advisory rule-based expert system for integrated pest management in Solanaceous crop systems. *Agricultural Systems, 76*(3), 1119–1135.
Manning, S., & Dix, A. (2008). Identifying students' mathematical skills from a multiple-choice diagnostic test using an iterative technique to minimise false positives. *Computers & Education, 51*(3), 1154–1171.
Medsker, L., Tan, M., & Turban, E. (1995). Knowledge acquisition from multiple experts: problems and issues. *Expert Systems with Applications, 9*(1), 40–55.
Mittal, S., & Dym, C. L. (1985). Knowledge acquisition from multiple experts. *AI Magazine, 6*(2), 32–36.
Nevin, A., Thousand, J., & Villa, R. (2007). Collaborative teaching: critique of the scientific evidence. In L. Florian (Ed.), *Handbook of special education research* (pp. 417–428). London, England: Sage Publishing.
Nevin, A., Thousand, J., & Villa, R. (2009). Collaborative teaching for teacher educators-what does the research say? *Teaching and Teacher Education, 25*(4), 569–574.
Offer, J., & Bos, B. (2009). The design and application of technology-based courses in the mathematics classroom. *Computers & Education, 53*(4), 1133–1137.
Pavlekovic, M., Zekic-Susac, M., & Djurdjevic, I. (2009). Comparison of intelligent systems in detecting a child's mathematical gift. *Computers & Education, 53*(1), 142–154.
Plessis, J. P. D., Biljon, J. A. V., Tolmie, C. J., & Wollinger, T. (1995). A model for intelligent computer-aided education systems. *Computers & Education, 24*(2), 89–106.
Springer, C. R., & Pear, J. J. (2008). Performance measures in courses using computer-aided personalized system of instruction. *Computers & Education, 51*(2), 829–835.
Stankov, S., Rosić, M., Žitko, B., & Grubišić, A. (2008). TEx-Sys model for building intelligent tutoring systems. *Computers & Education, 51*(3), 1017–1036.
Tsai, C. C., & Chou, C. (2002). Diagnosing students' alternative conceptions in science. *Journal of Computer Assisted Learning, 18*(2), 157–165.
Tseng, J. C. R., Chu, H. C., Hwang, G. J., & Tsai, C. C. (2008). Development of an adaptive learning system with two sources of personalization information. *Computers & Education, 51*(2), 776–786.
Tseng, S. S., Sue, P. C., Su, J. M., Weng, J. F., & Tsai, W. N. (2007). A new approach for constructing the concept map. *Computers & Education, 49*(3), 691–707.
Tung, M. C., Huang, J. Y., Keh, H. C., & Wai, S. S. (2009). Distance learning in advanced military education: analysis of joint operations course in the Taiwan military. *Computers & Education, 53*(3), 653–666.
Villa, R., Thousand, J., & Nevin, A. (2008). *A guide to co-teaching: Practical tips for facilitating student learning* (2nd ed.). Thousand Oaks, CA: Corwin Press.
Yang, Y. T. C., & Chan, C. Y. (2008). Comprehensive evaluation criteria for English learning websites using expert validity surveys. *Computers & Education, 51*(1), 403–422.
Zhou, Z. H., Jiang, Y., Yang, Y. B., & Chen, S. F. (2002). Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine, 24*(1), 25–36.