Expert Systems with Applications 39 (2012) 8380–8388

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A majority-density approach to developing testing and diagnostic systems with the cooperation of multiple experts based on an enhanced concept–effect relationship model

Dechawut Wanichsan [a], Patcharin Panjaburee [a,*], Parames Laosinchai [a], Wannapong Triampo [a,b], Sasithorn Chookaew [a]

[a] Institute for Innovative Learning, Mahidol University, 999, Phuttamonthon 4 Road, Salaya, Nakorn Pathom 73170, Thailand
[b] Department of Physics, Faculty of Science, Mahidol University, Rama VI, Bangkok 10400, Thailand

ARTICLE INFO

Keywords:
Concept–effect relationship model
Computer-based testing
Computer-assisted learning
Diagnostic learning system
Multi-expert system

ABSTRACT

In the recent years, diagnosing students' learning problems after testing and providing learning suggestions for them are an important research issue. Many studies have been conducted to develop a method for analyzing learning barriers of students such that helpful learning suggestions or guidance can be provided based on the analysis results. In this paper, we present a new procedure for integrating test item–concept relationship opinions based on majority density of multiple experts in order to enhance a concept–effect relationship model used for generating personalized feedback. It provides a useful and practical way to decrease inconsistencies in the weighting criteria of multiple experts and to enhance the entire learning-diagnosis procedure for developing testing and diagnostic systems.

1. Introduction

In conventional testing systems, a student is given a total score or grade as a test result to represent his/her learning status. Such feedback is insufficient to improve learning performance of students unless further guidance is also provided (Gerber, Grund, & Grote, 2008). This implies that diagnosing students' learning problems after testing and providing learning suggestions for them are an important research issue. During the past decade, many studies have been conducted to develop an effective method for analyzing learning barriers of students such that helpful learning suggestions or guidance can be provided based on the analysis results (Chen & Bai, 2009; Hwang, 2003, 2007). In the meantime, researchers have developed various computer-assisted testing and diagnostic systems for diagnosing students' learning problems and providing appropriate learning guidance for individual students on the Internet (Casamayor, Amandi, & Campo, 2009; Chen & Bai, 2009; Chiou, Hwang, & Tseng, 2009; Hwang, 2003; Sieber, 2009). For example, Chen (2008) developed a genetic-algorithm-based personalized learning system, in which a genetic algorithm was employed to generate appropriate learning paths based on the incorrect answers given by individual learners.

Panjaburee, Hwang, Triampo, and Shih (2010) presented a set of rules for integrating the weights of the relationship between a test item and a concept from multiple experts based on the concept–effect relationship (CER) model. Based on their method, a testing and diagnostic system was developed to detect students' learning problems and generate personalized feedback based on the relationship between prior and posterior knowledge while planning personalized learning paths for students. From the past experience, however, there were some drawbacks when applying Panjaburee et al.'s (2010) rules to develop testing and diagnostic systems because, for a single relationship between a test item and a concept, there were more than one rules used for integrating the weight values given by multiple experts when different weight values exist, because their rules did not consider the majority opinion from multiple experts, and because the confident level in making the decision was not considered during the integration of the weighting values. As a result, unreliable and low quality integrated weight values could be generated, resulting in equally unreliable and low quality learning suggestions given to the students.

To cope with these problems, we present an enhanced method for integrating weights of associated concepts for each test item from multiple experts. It provides a practical way to decrease inconsistencies in the weighting criteria of multiple experts and to enhance the entire learning-diagnosis procedure based on the CER model. The resultant testing and diagnostic systems should be able to provide reliable and high quality personalized learning guidance to students.

* Corresponding author.
E-mail addresses: kook260g@hotmail.com (D. Wanichsan), panjaburee_p@hotmail.com (P. Panjaburee), pl_one@hotmail.com (P. Laosinchai), scwtr@mahidol.ac.th (W. Triampo), nangfa1_edt@hotmail.com (S. Chookaew).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.182

The rest of this paper is organized as follows. In Section2, we briefly review the background of the CER model. In Section3, we briefly review Panjaburee et al.'s (2010) method for integrating the weights of associated concepts for each test item from multiple experts and some drawbacks. In Section4, we present a new method of integrating the weights for each test item with respect to concepts based on the majority opinion from multiple experts. In Section5, we use an example to show the procedure for integrating such weight values based on our new method, followed by a set of experiments in Section6. The conclusions are discussed in Section7.

2. The concept–effect relationship (CER) model

The CER model proposed by Hwang (2003) represents the prerequisite relationships among concepts that need to be learned in a specific order. Consider two concepts to be learned, say C_i and C_j. If C_i is a prerequisite to effectively understanding the more complex and higher level concept C_j, then the concept–effect relationship $C_i \rightarrow C_j$ is said to exist. For example, in computer programming, to learn the concept "Array", one might first need to learn "Variable and Data Type", while learning "Function" might require first learning both "Variable and Data Type" and "Expression and Operation". Fig. 1 presents an illustrative example of concept–effect relationships, which are important in diagnosing students' learning problems. For example, if a student fails to answer most of the test items concerning "Function", the problem is likely that the student has not thoroughly learned "Function" or its prerequisite concepts (such as "Variable and Data Type" or "Expression and Operation").

This model considers the relationship between prior and posterior knowledge while planning personalized learning paths. In this model, to provide learning suggestions to individual students, the error ratio (ER) for each student in answering the test items related to each concept needs to be analyzed; therefore, it is necessary to set the weight of association between test item Q_j and concept C_k (Hwang, 2003) in a Test Item Relationship Table (TIRT). Table 1 shows an example of a TIRT comprising four concepts and ten test items, where the TIRT (Q_j, C_k) is a value ranging from 0 to 5; "5" represents "high relevance" and "0" represents "no relevance". The error ratio (ER) for a student regarding concept C_k is then calculated by dividing the sum of TIRT (Q_j, C_k) values of the test items that the student failed to correctly answer by that of all of the test items. For example, assuming that a student failed to correctly answer Q_2, Q_6, and Q_{10}, we have $ER(C_1) = (2 + 1 + 1)/16 = 0.25$ and $ER(C_2) = (3 + 0 + 0)/9 = 0.33$.

As shown in Table 1, the values of ER's for a student to answer the test items concerning C_1, C_2, C_3, and C_4 are 0.25, 0.33, 0.50, and 0.60 respectively. We have

PATH1: C_1 (0.25) → C_2 (0.33),
PATH2: C_1 (0.25) → C_3 (0.50) → C_4 (0.60), and
PATH3: C_1 (0.25) → C_4 (0.60).

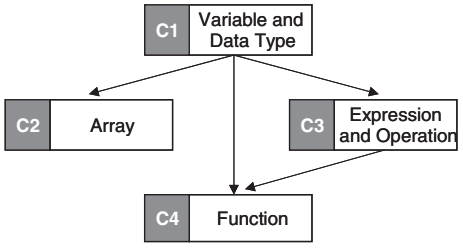

Fig. 1. An illustrative example of concept–effect relationships.

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

Test item	Concept, C_k			
	C_1	C_2	C_3	C_4
Q_1	2	0	0	0
Q_2	2	3	0	1
Q_3	4	0	0	0
Q_4	1	4	0	0
Q_5	1	0	2	2
Q_6	1	0	5	2
Q_7	1	2	1	0
Q_8	2	0	2	2
Q_9	1	0	2	0
Q_{10}	1	0	2	3
SUM(C_k)	16	9	14	10
ERROR(C_k)	4	3	7	6
ER(C_k)	0.25	0.33	0.50	0.60

A threshold θ is used to determine the acceptable error ratio. If $ER(C_k) < \theta$, the student is said to have learned concept C_k; 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 0.4, the poorly-learned paths are as follows:

PATH2: C_3 (0.50) → C_4 (0.60) and
PATH3: C_4 (0.60).

Therefore, the learning problems of the student could be a misunderstanding of concepts C_3 and C_4; moreover, the student should learn C_3 before learning C_4.

That is, in the existing model, the quality of the learning paths given to the students highly depends on the weight of association between test item Q_j and concept C_k given by the domain expert; therefore, subjective opinion, ignorance, or insufficient knowledge could affect the quality of the learning paths (Panjaburee et al., 2010). Such unreliable or low quality learning paths may be generated because the knowledge is usually acquired from a single expert.

Consider the example shown in Fig. 2; in the same test sheet, there are two domain experts, i.e., experts *A* and *B*, having different

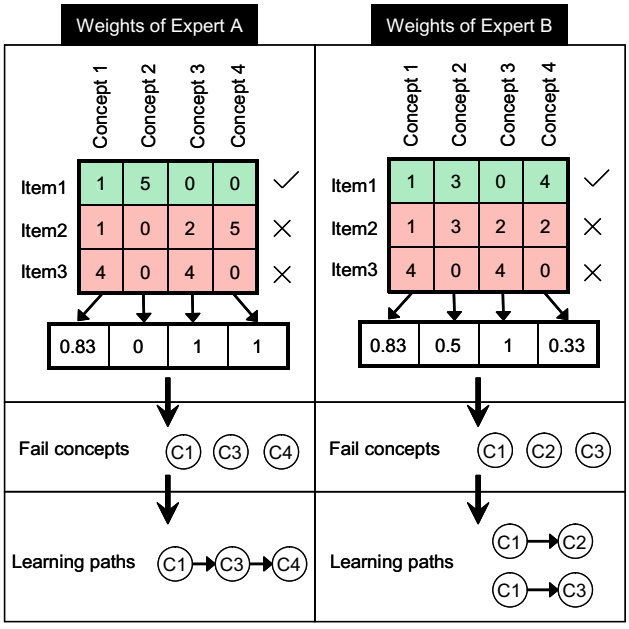

Fig. 2. Different opinions from two experts causing different learning paths.

opinions about the relationships among test items 1 and 2 and concepts 2 and 4. Let us assume that a student answers test items 2 and 3 incorrectly. In this case, the domain experts have different opinions about the weights to be taken into account; that is, the different ER's among concepts lead to different learning paths given to the student.

Panjaburee et al. (2010) proposed a set of rules to check and integrate the weights given by multiple experts to gain a consensus weight of association between test item Q_j and concept C_k, resulting in a consensus on the learning paths for students. However, there are some drawbacks of Panjaburee et al.'s (2010) method when applied to develop a testing and diagnostic system. In the next section, we will briefly review Panjaburee et al.'s method (2010) for integrating corresponding weights and its drawbacks.

3. A review of Panjaburee et al.'s method for integrating the weights of associated concepts for each test item

Panjaburee et al. (2010) presented rules to integrate the corresponding ⟨test item, concept⟩ relationships given by multiple experts based on the cooperation of experts. The following were conditions of Panjaburee et al. (2010) for integrating corresponding ⟨test item, concept⟩ relationships:

Step 1: Collect the ⟨test item, concept⟩ relationships of the test sheet from individual experts. An integer ranging from 1 to 5 (maximum weight) was used for representing a weighting value corresponding to "very weak", "weak", "average", "strong", and "very strong" relationships respectively. Moreover, "X" was used to represent "no relationship". In addition, the weight for the relationship between test item Q_j and concept C_k was represented by Weighting (Q_j, C_k), and the confident level for giving the weight was represented by Certainty (Q_j, C_k). The value of Certainty (Q_j, C_k) could be either "S" or "N", where "S" represented "Sure" for giving the weight, while "N" represented "Not sure". Assuming that n experts participated in the ⟨test item, concept⟩ determination process, the weight given by Expert E_i for test item Q_j and concept C_k was represented by Weighting (E_i, Q_j, C_k), and the confident level for Expert E_i in giving that value was represented by Certainty (E_i, Q_j, C_k). In this step, each expert was asked to provide the weights between the test items and the concepts. Let us assume that the test sheet, which covered four concepts, contained 10 items. Expert A's opinions for determining the weighting values in this test sheet were shown in Table 2, where C_1, C_2, C_3, and C_4 represented "Variable and Data Type", "Array", "Expression and Operation", and "Function" respectively.

Step 2: Integrate the corresponding ⟨test item, concept⟩ weights. While interpreting the corresponding weights, they called the values that were less than 3 the "weak side", and those that were greater than 3 the "strong side". They presented four categories, totaling fourteen rules (some of which are shown in Table 3), 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.

This step was repeated until no further checking and reconsidering weighting information was need.

Step 3: The integrated weights were used to construct TIRT (as shown in Table 1) from which ER could be calculated in order to diagnose learning problems of each student and generate a student's learning paths. The method assumed that high quality integrated weights generate high quality ER's; therefore reliable/high quality students' learning paths could be generated. That is, one way to enhance the entire learning diagnosis procedure based on CER model was to focus on integrating the corresponding ⟨test-item, concept⟩ weights.

Although Panjaburee et al. (2010) attempted to present a set of rules to check and integrate the weights of association between test item Q_j and concept C_k given by multiple experts in order to generate high quality weights, from the past experience, there were some drawbacks when applying Panjaburee et al.'s (2010) rules to develop testing and diagnostic systems. Because, for a single relationship between a test item and a concept, there could be more than one rules applicable to integrating the weighting values when different weight values existed, because their rules did not consider the majority opinion from multiple experts, and because the degree of confidence in making the decision was not considered during the integration of the weighting values, unreliable and low quality integrated weights could be generated; therefore, unreliable or low quality learning suggestions could be given to the students. In the next section, we will propose a new procedure for integrating the weights of ⟨test item, concept⟩ relationships to overcome the drawbacks of Panjaburee et al.'s (2010) set of rules.

Table 2
An illustrative example of the ⟨test item, concept⟩ relationships provided by a single expert.

Concept, C_k	Test, Q									
	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

4. A new procedure for integrating the weights of associated test items for each concept

In this section, we present a new procedure for integrating the weights of associated test items for each concept in testing and diagnostic systems based on the CER model. The set of weighting values and confident levels of associated test item Q_j for each concept C_k given by n experts are defined as $W_{Q_jC_k} = \{W_{Q_jC_k}(E_i) | i = 1 \text{ to } n\}$ and $CD_{Q_jC_k} = \{CD_{Q_jC_k}(E_i) | i = 1 \text{ to } n\}$ respectively, where $W_{Q_jC_k}(E_i) \in \{0,1,2,3,4,5\}$ and $CD_{Q_jC_k}(E_i) \in \{S, N\}$. $W_{Q_jC_k}(E_i) = 0$ means that expert E_i has determined that there is no relationship between test item Q_j and concept C_k, while $W_{Qj,Ck}(E_i) = 5$ means that expert E_i has determined that test item Q_j has very strong relationship with concept C_k. In addition, $CD_{Q_jC_k}(E_i) = \text{S}$ means that expert E_i has high confidence in determining the association between test item Q_j and concept C_k, whereas $CD_{Q_jC_k}(E_i) = \text{N}$ means that expert E_i has low confidence in determining such an association. The new procedure for integrating corresponding weighting values is presented as follows:

Step 1: Based on the weighting value $W_{Q_jC_k}(E_i)$ and the degree of confidence $CD_{Q_jC_k}(E_i)$, adjust the weighting value of each expert. The adjusted weighting value for test item Q_j and concept C_k of expert E_i is denoted by $adjW_{Q_jC_k}(E_i)$. While adjusting the weighting values, we shall call the values that are equal to or less than 2 the "weak side" and those greater than or equal to 3 the "strong side". Two conditions for adjusting the weighting values are as follows:

Condition 1

IF $W_{Q_jC_k}(E_i) \leqslant 2$ **AND** $CD_{Q_jC_k}(E_i), = \text{“N”}$,
THEN $adjW_{Q_jC_k}(E_i) = W_{Q_jC_k}(E_i) + 0.5$

Condition 1 is used for handling the case that a single expert has determined a weak side value with *low* confidence. In this case, the weighting value given by this expert is increased by 0.5. For

Table 3
An example of the knowledge-integration rules of Panjaburee et al. (2010).

Rule#	Condition	Integrated weight	Certainty level
7	There are some experts who assign the weight "X" with *high confidence*, moreover, there are some experts who assign *weak side* values or *strong side* values with *high confidence*	Reconsidering weights	–
8A	There are some experts who assign the weight "X" with *low confidence*, moreover, there are some experts who assign *weak side* values with *high confidence*	The minimum of *weak side* weights	"S"
8B	There are some experts who assign the weight "X" with *low confidence*, moreover, there are some experts who assign *strong side* values with *high confidence*	The minimum of *strong side* weights	"S"
9A	There are some experts who assign the weight "X" with *high confidence*, moreover, there are some experts who assign *weak side* values with *low confidence*	"X"	"S"
9B	There are some experts who assign the weight "X" with *high confidence*, moreover, there are some experts who assign *strong side* values with *low confidence*	"X"	"N"
11	There are some experts giving *weak side* values with *high confidence*, moreover, there are some experts giving *strong side* values with *high confidence*	Reconsidering weights	–
13	There is no expert giving "X", and there are some experts giving *weak side* values with *low confidence*, moreover, there are some experts giving *strong side* values with *high confidence*	The minimum of *strong side* weights	"S"
14	There is no expert giving "X", and there are some experts giving *weak side* values with *high confidence*, moreover, there are some experts giving *strong side* values with *low confidence*	The minimum of *weak side* weights	"S"

example, if expert E_1 has determined the association between test item Q_1 and concept C_2 as 0 with *low* confidence, this weighting value given by expert E_1 will be adjusted to 0.5.

Condition 2

IF $W_{Q_jC_k}(E_i) \geqslant 3$ **AND** $CD_{Q_jC_k}(E_i), =$ "N"
THEN $adjW_{Q_jC_k}(E_i) = W_{Q_jC_k}(E_i) - 0.5$

Condition 2 is used for handling the case that a single expert has determined a strong side value with *low* confidence. In this case, the weighting value given by this expert is decreased by 0.5. For example, if expert E_2 has determined the association between test item Q_2 and concept C_4 as 4 with *low* confidence, this weighting value given by expert E_2 will be adjusted to 3.5.

For a weighting value given by an expert with high confidence, that weighting value $W_{Q_jC_k}(E_i)$ is assigned to the value of $adjW_{Q_jC_k}(E_i)$. After adjusting all weighting values for test item Q_j and concept C_k, the set of those values of n experts is defined as $adjW_{Q_jC_k} = \{adjW_{Q_jC_k}(E_i) | i = 1 \text{ to } n\}$.

Step 2: Let $\max(W)$ and $\text{avg}(W)$ denote the maximum and average of weighting values in a set W, as defined by Eqs. (1) and (2). Based on a set of adjusted weighting values, $adjW_{Q_jC_k}$, calculate the density $dnstMX_{Q_jC_k}$ of all values around the maximum value in the set as shown in Eq. (3):

$$\max(W) = \max_{w \in W} w \tag{1}$$

$$\text{avg}(W) = \frac{1}{|W|} \sum_{w \in W} w \tag{2}$$

$$dnstMX_{Q_jC_k} = 1 - \frac{\max(adjW_{Q_jC_k}) - \text{avg}(adjW_{Q_jC_k})}{m} \tag{3}$$

where w represents a weighting value of an expert in the set W, $|W|$ denotes the cardinality of the set, and m represents the maximum rating scale (in this case, $m = 5$). Similarly, let $\min(W)$ denote the minimum of weighting values of experts, as defined by Eq. (4). Calculate the density $dnstMN_{Q_jC_k}$ of all values around the minimum as shown in Eq. (5):

$$\min(W) - \min_{w \in W} w \tag{4}$$

$$dnstMN_{Q_jC_k} = 1 - \frac{\text{avg}(adjW_{Q_jC_k}) - \min(adjW_{Q_jC_k})}{m} \tag{5}$$

The less the value of $dnstMX_{Q_jC_k}$ compared to that of $dnstMN_{Q_jC_k}$, the higher the probability that the majority of opinions from n experts are closer to the value of $\min(adjW_{Q_jC_k})$; that is, the value of $\max(adjW_{Q_jC_k})$ is defined as a potential *outlier*, which represents an extreme weighting value among values for test item Q_j and concept C_k given by n experts. The more the value of $dnstMX_{Q_jC_k}$ compared to that of $dnstMN_{Q_jC_k}$, the higher the probability that the majority of opinions from n experts are closer to the value of $\max(adjW_{Q_jC_k})$; that is, the value of $\min(adjW_{Q_jC_k})$ is defined as a potential *outlier*. If the value of $dnstMX_{Q_jC_k}$ is equal to that of $dnstMN_{Q_jC_k}$, there is no *outlier* among the weight values. The next step, Step 3, is required to verify whether the potential outlier should be removed. In the case of no *outlier*, Step 3 is unnecessary; therefore, all weighting values will be used to calculate an integrated weight in Step 4.

Step 3: Based on the potential outlier, verify whether it is distant enough from others and proper to be removed. Let $\text{MAD}(W)$ denote the mean absolute deviation of weighting values in a set W, as defined by Eq. (6). For test item Q_j and concept C_k, calculate the distance $distOW_{Q_jC_k}$ between the potential outlier and the average of the remaining adjusted weighting values as shown in Eq. (7):

$$\text{"MAD"}(W) = 1/|W| \sum_{w \in W} |w - \text{avg}(W)| \tag{6}$$

$$distOW_{Q_jC_k} = \left| outlier - \text{avg}(adjW_{Q_jC_k} - \{outlier\}) \right| \tag{7}$$

where *outlier* represents the extreme value detected in Step 2. The condition to consider whether the outlier should be removed is:

IF $distOW_{Q_jC_k} > 1.25$ **AND**
$distOW_{Q_jC_k} > \eta \times \text{MAD}(adjW_{Q_jC_k} - \{outlier\})$
THEN $adjW_{Q_jC_k} = adjW_{Q_jC_k} - \{outlier\}$
REPEAT Step 2
ELSE GOTO Step 4

Here, 1.25, which is one-fourth the maximum rating scale m (in this case, $m = 5$), is chosen so that any value within $\text{avg}(adjW_{Q_jC_k} - \{outlier\}) \pm 1.25$, which covers half the scale, would not be considered an outlier. This first condition is employed to prevent against false removal of the potential outlier by the second condition, which detects its relative distance from the rest of the weights. To that end, a symmetric triangular distribution is employed as the model of experts' behavior. Since sampling from the same population with smaller sample size results in smaller variation, this effect for the triangular distribution is estimated by averaging 10,000 mean absolute deviations (MAD) of independent samples

Table 4
The averaged MAD of continuous triangular distribution data and the value of η.

Sample size	Average MAD (10,000 rounds)	η = (0.5/average MAD)
2	0.117	4.29
3	0.135	3.70
4	0.143	3.49
5	0.148	3.37
6	0.152	3.30
7	0.153	3.26
8	0.155	3.23
9	0.157	3.19
10	0.158	3.16
11	0.160	3.13
12	0.160	3.13
13	0.160	3.13
14	0.160	3.13
15	0.160	3.13

for each sample size n. η then equals half the range of the triangular distribution divided by this average MAD. That is, $\eta \times \text{MAD}(adjW_{Q_jC_k} - \{outlier\})$ is the critical value beyond which a weight could be considered a true outlier. If both conditions are met, the outlier will be removed from the set of adjusted weighting values $adjW_{Q_jC_k}$. Moreover, to detect other outliers, Step 2 will be repeated once again. On the other hand, if the potential outlier is not suitable to be removed, all weighting values will be used to calculate an integrated weight in Step 4. Table 4 shows values of average MAD and values of η when $n = 2$ to 15.

Step 4: After removing outliers, the integrated adjusted weighting value $iWeight_{Q_jC_k}$ for test item Q_j and concept C_k is calculated as follows:

$$iWeight_{Q_jC_k} = \text{avg}(adjW_{Q_jC_k}) \tag{8}$$

To verify that the value of $iWeight_{Q_jC_k}$ is reasonable, the degree of majority opinion degree $dMO_{Q_jC_k}$ needs to be calculated as follows:

$$dMO_{Q_jC_k} = 1 - \frac{1}{m}\text{MAD}(adjW_{Q_jC_k}) \tag{9}$$

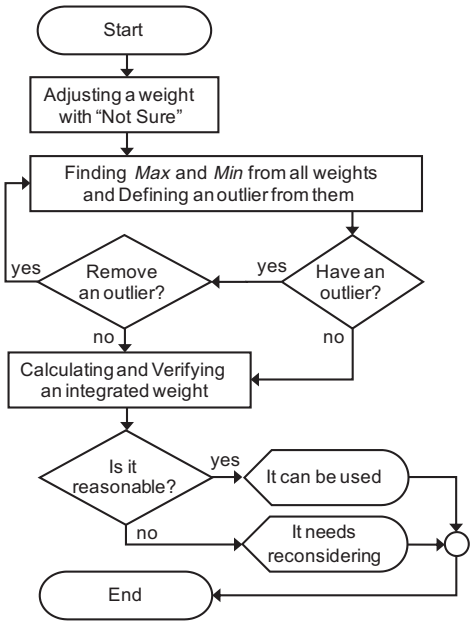

Fig. 3. A diagram describing the proposed method for integrating experts' weighting values.

where m represents the maximum rating scale (in this case, $m = 5$). The higher the value of $dMO_{Q_jC_k}$, the higher probability that the experts have agreed on the value of $iWeight_{Q_jC_k}$. A threshold, θ, is used to indicate the acceptable agreement level. In this case, the value of θ is assumed to be 0.85. If $dMO_{Q_jC_k} < \theta$, the experts will be asked to reconsider and discuss their weighting values; otherwise, the integrated weighting value will be used in any testing and diagnostic systems. A summary of our proposed method is shown in Fig. 3. For a test item and a concept, in the first step, weighting values with low confidence will be adjusted. Outliers will be then detected and they will be removed if necessary. Finally, an integrated weighting value will be calculated and verified whether it is reasonable or it needs to be reconsidered.

5. Examples

5.1. Example 5.1

Let us compare the weighting-value integration for test item Q_2 and concept C_4 from five experts using Panjaburee et al.'s (2010) rules with our new weighting-value integration procedure. Assume that the weighting values and the degrees of confidence for test item Q_2 and concept C_4 given by five experts are as follows:

	E_1	E_2	E_3	E_4	E_5
$W_{Q_2C_4}, CD_{Q_2C_4}$	2, S	2, N	3, S	4, N	4, S

Step 1: Based on the elicited weighting values from five experts, we can see that experts E_2 and E_4 have determined the weighting values for test item Q_2 and concept C_4 with low confidence. Therefore these weighting values are adjusted as follows:

	E_1	E_2	E_3	E_4	E_5
$adjW_{Q_2C_4}$	2	2.5	3	3.5	4

Step 2: Based on the values of $adjW_{Q_2C_4}$ given by, $\max adj(W_{Q_2C_4})$ is 4 and $\min(adjW_{Q_2C_4})$ is 2. Therefore the $dnstMX_{Q_2C_4}$ and $dnstMN_{Q_2C_4}$ values are evaluated as follows:

$$\text{avg}(\{2, 2.5, 3, 3.5, 4\}) = 3$$

$$dnstMX_{Q_2C_4} = 1 - \frac{4-3}{5} = 0.80$$

$$dnstMN_{Q_2C_4} = 1 - \frac{3-2}{5} = 0.80$$

We can see that $dnstMX_{Q_2C_4}$ is equal to $dnstMN_{Q_2C_4}$; that is, there is no outlier and the next step is Step 4 wherein the integrated weighting value will be calculated.

Step 4: The set of adjusted weighting values is {2, 2.5, 3, 3.5, 4}. Therefore, the integrated weighting value for test item Q_2 and concept C_4 is calculated by Eq. (8) as follows:

$$iWeight_{Q_2C_4} = \text{avg}(\{2, 2.5, 3, 3.5, 4\}) = 3$$

To verify that the value of $iWeight_{Q_2C_4}$ is reasonable, the value of $dMO_{Q_2C_4}$ is calculated using Eq. (9) as follows:

$$dMO_{Q_2C_4} = 1 - \frac{1}{5}\text{MAD}(\{2, 2.5, 3, 3.5, 4\}) = 0.88$$

We can see that the value of $dMO_{Q_2C_4}$ is greater than 0.85; that is, the experts have agreed on the value $iWeight_{Q_2C_4} = 3$, as the integrated weighting value.

It should be noted that the integrated weighting value "4" for test item Q_2 and concept C_4 using Rule 13 (Panjaburee et al., 2010) and "2" using Rule 14 (Panjaburee et al., 2010) are unreliable, which can be described as follows. From a set of weighting

values given by five experts for test item Q_2 and concept C_4, we can see that there are two rules, "Rule 13" and "Rule 14" that can be used for handling the set of weighting values. Interestingly, the integrated weighting values are the maximum and minimum of all weights, respectively. That is, an unreliable integrated weighting value results from using Panjaburee et al.'s (2010) rules because their rules did not consider the majority opinion from five experts. The integrated weighting value "3" calculated using the new weighting integration procedure is clearly more reliable, because this new procedure considers the majority opinion for giving the integrated weighting value. Therefore, the proposed procedure can overcome the drawbacks of the set of rules presented by Panjaburee et al. (2010) and more reliably calculate the integrated weighting value for diagnosing learning problems in testing and diagnostic systems based on the CER model.

5.2. Example 5.2

Let us compare the weighting-value integration for test item Q_3 and concept C_5 from six experts using Panjaburee et al.'s (2010) rules with our new weighting-value integration procedure. Assume that the weighting values and the degrees of confidence for test item Q_3 and concept C_5 given by six experts are as follows:

	E_1	E_2	E_3	E_4	E_5	E_6
$W_{Q_3C_5}, CD_{Q_3C_5}$	0, S	0, N	1, S	1, S	2, S	2, N

Step 1: Based on the elicited weighting values from six experts, we can see that experts E_2 and E_6 have determined the weighting values for test item Q_3 and concept C_5 with low confidence. Therefore these weighting values are adjusted as follows:

	E_1	E_2	E_3	E_4	E_5	E_6
$adjW_{Q_3C_5}$	0	0.5	1	1	2	2.5

Step 2: Based on the values of $adjW_{Q_3C_5}$ given by, $\max(adjW_{Q_3C_5})$ is 2.5 and $\min(adjW_{Q_3C_5})$ is 0. Therefore the $dnstMX_{Q_3C_5}$ and $dnstMN_{Q_3C_5}$ values are evaluated as follows:

$$\text{avg}(\{0, 0.5, 1, 1, 2, 2.5\}) = 7/6$$
$$dnstMX_{Q_3C_5} = 1 - \frac{2.5 - 7/6}{5} = 0.73$$
$$dnstMN_{Q_3C_5} = 1 - \frac{7/6 - 0}{5} = 0.77$$

We can see that $dnstMN_{Q_3C_5}$ is greater than $dnstMX_{Q_3C_5}$; that is, the majority opinion from six experts is closer to the value 0, and the value 2.5 is defined as a potential *outlier*. Consequently, Step 3 will be used to verify whether it is an outlier.

Step 3: Based on the value 2.5, verify whether it is distant enough from others and proper to be removed. $distOW_{Q_3C_5}$ and $\text{MAD}(adjW_{Q_3C_5} - \{outlier\})$ are calculated.

$$distOW_{Q_3C_5} = |2.5 - \text{avg}(\{0, 0.5, 1, 1, 2\})| = 1.60$$
$$\eta \times \text{MAD}(\{0, 0.5, 1, 1, 2\}) = 3.37 \times 0.52 = 1.75$$

Because $distOW_{Q_3C_5}$ is more than 1.25 (1.60 > 1.25), but it is still less than $\eta \times \text{MAD}(adjW_{Q_3C_5} - \{outlier\})$ (1.60 < 1.75), the value 2.5 will be not removed from the set $adjW_{Q_3C_5}$. Therefore, all weighting values will be used to calculate an integrated weighting value in Step 4.

Step 4: Because there is no outlier removed, the set of adjusted weighting values $adjW_{Q_3C_5}$ is {0, 0.5, 1, 1, 2, 2.5}. Therefore, the integrated weighting value for test item Q_3 and concept C_5 is calculated by Eq. (8) as follows:

$$iWeight_{Q_3C_5} = \text{avg}(\{0, 0.5, 1, 1, 2, 2.5\}) = 1.17$$

To verify that the value of $iWeight_{Q_3C_5}$ is reasonable, the value of $dMO_{Q_3C_5}$ is calculated using Eq. (9) as follows:

$$dMO_{Q_3C_5} = 1 - \frac{1}{5}\text{MAD}(\{0, 0.5, 1, 1, 2, 2.5\}) = 0.856$$

We can see that the value of $dMO_{Q_3C_5}$ is greater than 0.85; that is, the experts have agreed on the value 1.17 as the integrated weighting value.

It should be noted that the integrated weighting value "0" for test item Q_3 and concept C_5 using Rule 9A (Panjaburee et al., 2010) and "experts need to reconsider their weight" using Rule 7 (Panjaburee et al., 2010) are unreliable, which can be described as follows. From a set of weighting values given by six experts for test item Q_3 and concept C_5, we can see that there are two rules, "Rule 7" and "Rule 9A"that can used for handling the set of weighting values. Interestingly, the integrated weighting value "0" should not be used because most weighting values are more than "0". That is, there is an unreliable integrated weighting value while using the Panjaburee et al.'s (2010) rules because their rules did not consider the majority opinion from six experts. The integrated weighting value "1.17" calculated using the new weighting integration procedure is clearly more reliable, because this new procedure considers the majority opinion for giving the integrated weighting value. Therefore, the proposed procedure can overcome the drawbacks of the set of rules presented by Panjaburee et al. (2010) and more reliably calculate the integrated weighting value for diagnosing learning problems in testing and diagnostic systems based on the CER model.

5.3. Example 5.3

Let us compare the weighting-value integration for test item Q_1 and concept C_9 from seven experts using Panjaburee et al.'s (2010) rules with our new weighting-value integration procedure. Assume that the weighting values and the degrees of confidence for test item Q_1 and concept C_9 given by seven experts are as follows:

	E_1	E_2	E_3	E_4	E_5	E_6	E_7
$W_{Q_1C_9}, CD_{Q_1C_9}$	0, S	1, N	4, N	4, S	5, N	5, N	5, S

Step 1: Based on the elicited weighting values from seven experts, we can see that experts E_2, E_3, E_5 and E_6 have determined the weighting values for test item Q_1 and concept C_9 with low confidence. Therefore these weighting values are adjusted as follows:

	E_1	E_2	E_3	E_4	E_5	E_6	E_7
$adjW_{Q_1C_9}$	0	1.5	3.5	4	4.5	4.5	5

Step 2: Based on the values of $adjW_{Q_1C_9}$, $\max(adjW_{Q_1C_9})$ is 5 and $\min(adjW_{Q_1C_9})$ is 0. Therefore the $dnstMX_{Q_1C_9}$ and $dnstMN_{Q_1C_9}$ values are evaluated as follows:

$$\text{avg}(\{0, 1.5, 3.5, 4, 4.5, 4.5, 5\}) = 23/7$$
$$dnstMX_{Q_1C_9} = 1 - \frac{5 - 23/7}{5} = 0.66$$
$$dnstMN_{Q_1C_9} = 1 - \frac{23/7 - 0}{5} = 0.34$$

We can see that $dnstMX_{Q_1C_9}$ is greater than $dnstMN_{Q_1C_9}$; that is, the majority opinion from seven experts is closer to the value 5, and the value 0 is defined as a potential *outlier*. Consequently, the next step is Step 3 which verifies whether 0 is an outlier.

Step 3: Based on the value 0, verify whether it is distant enough from others and proper to be removed. $distOW_{Q_1C_9}$ and MAD $(adjW_{Q_1C_9} - \{outlier\})$ are calculated.

$$distOW_{Q_1C_9} = |0 - \text{avg}(\{1.5, 3.5, 4, 4.5, 4.5, 5\})| = 3.83$$
$$\eta \times \text{MAD}(\{1.5, 3.5, 4, 4.5, 4.5, 5\}) = 3.30 \times 0.89 = 2.93$$

Because $distOW_{Q_1C_9}$ is more than 1.25 (3.83 > 1.25) and $\eta \times$ MAD $(adjW_{Q_1C_9} - \{outlier\})$ (3.83 > 2.93), the value 0 is removed from the set $adjW_{Q_1C_9}$. Consequently, Step 2 is used once again to detect another outlier.

Step 2 (the 2nd round): After removing the outlier, the set $adjW_{Q_1C_9} = \{1.5, 3.5, 4, 4.5, 4.5, 5\}$.max$(adjW_{Q_1C_9})$ is 5 and min $(adjW_{Q_1C_9})$ is 1.5. Therefore the $dnstMX_{Q_1C_9}$ and $dnstMN_{Q_1C_9}$ values are evaluated as follows:

$$\text{avg}(\{1.5, 3.5, 4, 4.5, 4, 5, 5\}) = \frac{23}{6}$$
$$dnstMX_{Q_1C_9} = 1 - \frac{5 - 23/6}{5} = 0.77$$
$$dnstMN_{Q_1C_9} = 1 - \frac{23/6 - 1.5}{5} = 0.53$$

The value of $dnstMX_{Q_1C_9}$ is greater than $dnstMN_{Q_1C_9}$; that is, the majority opinion from six experts is closer to the value 5, and the value 1.5 is defined as a potential *outlier*. Consequently, Step 3 will be used to verify whether it is an outlier.

Step 3 (the 2nd round): Based on the value 1.5, verify whether it is distant enough from others and proper to be removed. $dnstOW_{Q_1C_9}$ and MAD$(adjW_{Q_1C_9} - \{outlier\})$ are calculated.

$$distOW_{Q_1C_9} = |1.5 - \text{avg}(\{3.5, 4, 4.5, 4.5, 5\})| = 2.80$$
$$\eta \times \text{MAD}(\{3.5, 4, 4.5, 4.5, 5\}) = 3.37 \times 0.44 = 1.48$$

Because $distOW_{Q_1C_9}$ is more than 1.25 (2.80 > 1.25) and $\eta \times$ MAD $(adjW_{Q_1C_9} - \{outlier\})$ (2.80 > 1.48), the outlier (1.5) is removed from the set $adjW_{Q_1C_9}$. Consequently, Step 2 is used again to detect another outlier.

Step 2 (the 3rd round): After removing the outlier, the set $adjW_{Q_1C_9} = \{3.5, 4, 4.5, 4.5, 5\}$. max$(adjW_{Q_1C_9})$ is 5 and min $(adjW_{Q_1C_9})$ is 3.5. Therefore the $dnstMX_{Q_1C_9}$ and $dnstMN_{Q_1C_9}$ values are evaluated as follows:

$$\text{avg}(\{3.5, 4, 4.5, 4, 5, 5\}) = 4.3$$
$$dnstMX_{Q_1C_9} = 1 - \frac{5 - 4.3}{5} = 0.86$$
$$dnstMN_{Q_1C_9} = 1 - \frac{4.3 - 3.5}{5} = 0.84$$

The value of $dnstMX_{Q_1C_9}$ is greater than $dnstMN_{Q_1C_9}$; that is, the majority opinion from five experts is closer to the value 5, and the value 3.5 is defined as a potential *outlier*. Consequently, Step 3 is used to verify such outlier once again.

Step 3 (the 3rd round): Based on the value 3.5, verify whether it is distant enough from others and proper to be removed. $distOW_{Q_1C_9}$ and MAD$(adjW_{Q_1C_9} - \{outlier\})$ are calculated.

$$distOW_{Q_1C_9} = |3.5 - \text{avg}(\{4, 4.5, 4.5, 5|)| = 1$$
$$\eta \times \text{MAD}(\{4, 4.5, 4.5, 5\}) = 3.49 \times 0.25 = 0.87$$

Because $distOW_{Q_1C_9}$ is less than 1.25 (1 < 1.25)), the value 3.5 is not removed from the set $adjW_{Q_1C_9}$. The remaining weighting values will be used to calculate an integrated weighting value in Step 4.

Step 4: After removing outliers, the set of adjusted weighting values $adjW_{Q_1C_9}$ is {3.5, 4, 4.5, 4.5, 5}. Therefore, the integrated weighting value for test item Q_1 and concept C_9 is calculated by Eq. (8) as follows:

$$iWeight_{Q_1C_9} = \text{avg}(\{3.5, 4, 4.5, 4.5, 5\}) = 4.3$$

To verify that the value of $iWeight_{Q_1C_9}$ is reasonable, the value of $dMO_{Q_1C_9}$ is calculated using Eq. (9) as follows:

$$dMO_{Q_1C_9} = 1 - \frac{1}{5}\text{MAD}(\{3.5, 4, 4.5, 4.5, 5\}) = 0.91$$

We can see that the value of $dMO_{Q_1C_9}$ is greater than 0.85; that is, the experts have agreed on the value 4.3 as the integrated weighting value.

It should be noted that the integrated weighting value "0" for test item Q_1 and concept C_9 using Rule 9A (Panjaburee et al., 2010) and "experts need to reconsider their weight" using Rule 7 (Panjaburee et al., 2010) are unreliable, which can be described as follows. From a set of weighting values given by seven experts for test item Q_1 and concept C_9, we can see that there are two rules, "Rule 7" and "Rule 9A" that can be used for handling the set of weighting values. Interestingly, we can see that the integrated weighting value "0" calculated by "Rule 9A" is in the weak side and almost all weighting values are in the strong side. That is, there is an unreliable integrated weighting value while using the Panjaburee et al.'s (2010) rules because their rules did not consider the majority opinion from seven experts. The integrated weighting value "4.3" calculated using the new weighting integration procedure is clearly more reliable, because this new procedure considers the majority opinion for giving the integrated weighting value. Therefore, the proposed procedure can overcome the drawbacks of the set of rules presented in Panjaburee et al. (2010) and more reliably calculate the integrated weighting value for diagnosing learning problems in testing and diagnostic systems based on the CER model.

5.4. Example 5.4

Let us compare the weighting-value integration for test item Q_6 and concept C_7 from eight experts using Panjaburee et al.'s (2010) rules with our new weighting-value integration procedure. Assume that the weighting values and the degrees of confidence for test item Q_6 and concept C_7 given by eight experts are as follows:

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8
$W_{Q_6C_7}, CD_{Q_6C_7}$	0, S	0, S	0, N	1, S	5, N	5, S	5, S	5, S

Step 1: Based on the elicited weighting values from eight experts, we can see that experts E_3 and E_5 have determined the weighting values for test item Q_6 and concept C_7 with low confidence. Therefore these weighting values are adjusted as follows:

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8
$adjW_{Q_6C_7}$	0	0	0.5	1	4.5	5	5	5

Step 2: Based on the values of $adjW_{Q_6C_7}$ given by, max$(adjW_{Q_6C_7})$ is 5 and min$(adjW_{Q_6C_7})$ is 0. Therefore the $dnstMX_{Q_6C_7}$ and $dnstMN_{Q_6C_7}$ values are evaluated as follows:

$$\text{avg}(\{0, 0, 0.5, 1, 4.5, 5, 5, 5\}) = 21/8$$
$$dnstMX_{Q_6C_7} = 1 - \frac{5 - 21/8}{5} = 0.53$$
$$dnstMN_{Q_6C_7} = 1 - \frac{21/8 - 0}{5} = 0.48$$

we can see that $dnstMX_{Q_6C_7}$ is greater than $dnstMN_{Q_6C_7}$; that is, the majority opinion from eight experts is closer to the value 5, and the value 0 is defined as a potential *outlier*. Consequently, Step 3 is used to verify whether it is an outlier.

Table 5
The parameters used to randomize dataset.

Pattern	Random weighting values with triangular distribution			Random confidence degree with uniform distribution	
	Lower limit (a)	Upper limit (b)	Mode (c)	Lower limit	Upper limit
1	0	5	0	0	1
2	0	5	2.5	0	1
3	0	5	5	0	1

Table 6
The comparison results of the number of weighting value reconsiderations during the integration of associated test items for each concept from multiple experts.

Pattern	Data set	# of expert	# of item	# of concept	# of reconsiderations	
					Panjaburee et al. (2010) (%)	The proposed method (%)
1	1	3	30	5	34.4	10.4
	2	4	30	5	50.1	26.8
	3	5	30	5	62.5	35.9
	4	6	30	5	71.6	42.4
	5	7	30	5	78.4	47.0
	6	8	30	5	83.3	51.0
	7	9	30	5	87.0	54.1
	8	10	30	5	90.0	56.7
2	1	3	30	5	29.8	9.1
	2	4	30	5	45.8	21.9
	3	5	30	5	58.7	28.6
	4	6	30	5	68.1	33.3
	5	7	30	5	75.2	36.9
	6	8	30	5	80.3	40.2
	7	9	30	5	84.2	42.7
	8	10	30	5	87.0	44.7
3	1	3	30	5	32.4	10.2
	2	4	30	5	48.9	26.7
	3	5	30	5	61.4	36.1
	4	6	30	5	71.4	42.6
	5	7	30	5	78.1	46.8
	6	8	30	5	83.0	51.2
	7	9	30	5	86.5	54.3
	8	10	30	5	89.2	57.0

Step 3: Based on the value 0, verify whether it is distant enough from others and proper to be removed. $distOW_{Q_6C_7}$ and $\mathrm{MAD}(adjW_{Q_6C_7} - \{outlier\})$ are calculated.

$$distOW_{Q_6C_7} = |0 - \mathrm{avg}(\{0, 0.5, 1, 4.5, 5, 5, 5\})| = 3$$
$$\eta \times \mathrm{MAD}(\{0, 0.5, 1, 4.5, 5, 5, 5\}) = 3.26 \times 2.14 = 6.99$$

Because $distOW_{Q_6C_7}$ is more than 1.25 ($3 > 1.25$), and but it is still less than $\eta \times \mathrm{MAD}(adjW_{Q_6C_7} - \{outlier\})$ ($3 < 6.99$), the outlier (0) is not removed from the set $adjW_{Q_6C_7}$. All weighting values in the set will be used for calculating an integrated weighting value in Step 4.

Step 4: Because no outlier is removed, the set of adjusted weighting values $adjW_{Q_6C_7}$ is {0, 0, 0.5, 1, 4.5, 5, 5, 5}. Therefore, the integrated weighting value for test item Q_6 and concept C_7 is calculated by Eq. (8) as follows:

$$iWeight_{Q_6C_7} = \mathrm{avg}(\{0, 0, 0.5, 1, 4.5, 5, 5, 5\}) = 2.63$$

To verify whether the value of $iWeight_{Q_6C_7}$ is reasonable, the value of $dMO_{Q_6C_7}$ is calculated using Eq. (9) as follows:

$$dMO_{Q_6C_7} = 1 - \frac{1}{5}\mathrm{MAD}(\{0, 0, 0.5, 1, 4.5, 5, 5, 5\}) = 0.55$$

We can see that the value of $dMO_{Q_6C_7}$ is not greater than 0.85; that is, the experts need to check, discuss and reconsider their weighting value.

It should be noted that the integrated weighting value "5" for test item Q_6 and concept C_7 using Rule 8B (Panjaburee et al., 2010), the value "0" using Rule 9B (Panjaburee et al., 2010), and "experts need to reconsider their value" using Rule 7 (Panjaburee et al., 2010) are unreliable, which can be described as follows. From a set of weighting values given by eight experts for test item Q_6 and concept C_7, we can see that there are three rules, "Rule 7", "Rule 8B" and "Rule 9B", that can be used for handling the set of weighting values. Because opinions of experts are separated into two groups, one on the weak side and the other on the strong side, clearly, the result "experts need to reconsider their value" determined by using the new weighting integration procedure is reasonable.

6. Experimental results

Let us compare the number of reconsiderations of weighting values resulting from the set of rules in Panjaburee et al. (2010) to that resulting from the proposed procedure using random data sets. Weighting values were generated from the triangular distribution while degrees of confidence were generated from the uniform distribution. To reduce bias of this experiment, we utilized three modes, 0, 2.5, and 5, to represent three patterns of the triangular-distribution datasets; the parameters used for generating artificial datasets were shown in Table 5. Moreover, the experiment for each dataset was repeated 1000 times. Table 6 shows the percentage of reconsideration of the weighting values generated by the set of rules in Panjaburee et al. (2010) and the proposed procedure, averaged over 1000 runs. When the number of experts who participate in determining the association between test items

and concepts was increased, clearly, our proposed method resulted in a much lower number of reconsiderations on every dataset in every pattern. From these results, it implies that the consideration of majority opinion in the proposed procedure is an important factor in decreasing the number of reconsiderations of the weighting values. Therefore, the proposed procedure can overcome the drawbacks of the set of rules presented in Panjaburee et al. (2010) and reduce the amount of time used for discussing, checking, and reconsidering the weighting values when conflicting opinions from multiple experts exist.

7. Conclusions

The integration of weighting values given to the associated test item for each concept from multiple experts is an important issue for developing testing and diagnostic systems based on the CER model. In this paper, we presented a new procedure for integrating weighting values of the associated test item for each concept from multiple experts. The proposed procedure considers the degree of confidence in making the decision for the weighting value, the majority opinion, and the reliability of the integrated weighting value. It provides a useful way to integrate the weighting values while developing testing and diagnostic systems based on the CER model. It can overcome the drawbacks of the set of rules presented in Panjaburee et al. (2010), resulting in more reliable or better quality integrated weighting values, and enhance the entire learning diagnosis procedure based on the CER model.

Acknowledgment

This work was supported in part by the Mahidol University (MU) research fund.

References

Casamayor, A., Amandi, A., & Campo, M. (2009). Intelligent assistance for teachers in collaborative e-learning environments. *Computers and Education, 53*(4), 1147–1154.

Chen, C.-M. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers and Education, 51*(2), 787–814.

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.

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 and Education, 53*(2), 261–272.

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.

Hwang, G.-J. (2003). A conceptual map model for developing intelligent tutoring systems. *Computers and Education, 40*(3), 217–235.

Hwang, G.-J. (2007). Gray forecast approach for developing distance learning and diagnostic systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37*(1), 98–108.

Panjaburee, P., Hwang, G.-J., Triampo, W., & Shih, B.-Y. (2010). A multi-expert approach for developing testing and diagnostic systems based on the concept–effect model. *Computers and Education, 55*(2), 527–540.

Sieber, V. (2009). Diagnostic online assessment of basic IT skills in 1st-year undergraduates in the Medical Sciences Division, University of Oxford. *British Journal of Educational Technology, 40*(2), 215–226.