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

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.

Article Info

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.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.

do: 10.1016/j.eswa.2012.01.182

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.

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).
The rest of this paper is organized as follows. In Section 2, we briefly review the background of the CER model. In Section 3, 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 Section 4, 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 Section 5, 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 Section 6. The conclusions are discussed in Section 7.

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 C1 and C2. If C1 is a prerequisite to effectively understanding the more complex and higher level concept C2, then the concept–effect relationship C1 → C2 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 Qi and concept Ck (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 (Qi, Ck) 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 Ck is then calculated by dividing the sum of TIRT (Qi, Ck) values of the test items that the student failed to correctly answer by that of all the test items. For example, assuming that a student failed to correctly answer Q2, Q6, and Q10, we have ER(C1) = (2 + 1 + 1)/16 = 0.25 and ER(C4) = (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 C1, C2, C3, and C4 are 0.25, 0.33, 0.50, and 0.60 respectively. We have

PATH1: C1 (0.25) → C2 (0.33),
PATH2: C1 (0.25) → C4 (0.60), and
PATH3: C1 (0.25) → C4 (0.60).

A threshold θ is used to determine the acceptable error ratio. If ER(Ck) < θ, the student is said to have learned concept Ck; 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: C1 (0.50) → C4 (0.60) and
PATH3: C1 (0.60).

Therefore, the learning problems of the student could be a misunderstanding of concepts C1 and C4; moreover, the student should learn C1 before learning C4.

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 Qj and concept Ck 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 expert weights for the same problem. Therefore, different learning paths are derived. The following figure shows two learning paths generated by experts A and B.
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 and concept Ck, 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 and concept Ck was represented by Weighting (Q, Ck), and the confident level for giving the weight was represented by Certainty (Q, Ck). The value of Certainty (Q, Ck) 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 Ei for test item Q and concept Ck was represented by Weighting (Ei, Q, Ck), and the confident level for Expert Ei in giving that value was represented by Certainty (Ei, Q, Ck). 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 test items. Expert Ei’s opinions about the weights to be taken into account; that is, the domain experts have different conditions of Panjaburee et al. (2010) for integrating corresponding (test item, concept) relationships to overcome the drawbacks of Panjaburee et al.’s (2010) set of rules.

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 for each concept Ck given by n experts are defined as WQjCk = {WQjCk(Ei)|i = 1 to n} and CDQjCk = {CDQjCk(Ei)|i = 1 to n} respectively, where WQjCk(Ei) ∈ {0, 1, 2, 3, 4, 5} and CDQjCk(Ei) ∈ {S, N}. WQjCk(Ei) = 0 means that expert Ei has determined that there is no relationship between test item Q and concept Ck, while WQjCk(Ei) ≥ 5 means that expert Ei has determined that test item Q has very strong relationship with concept Ck. In addition, CDQjCk(Ei) = S means that expert Ei has high confidence in determining the association between test item Q and concept Ck, whereas CDQjCk(Ei) = N means that expert Ei 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 WQjCk(Ei) and the degree of confidence CDQjCk(Ei), adjust the weighting value of each expert. The adjusted weighting value for test item Q and concept Ck of expert Ei is denoted by adjWQjCk(Ei). 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 WQjCk(Ei) ≤ 2 AND CDQjCk(Ei) = “N”,

THEN adjWQjCk(Ei) = WQjCk(Ei) + 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
example, if expert $E_1$ has determined the association between test item $Q_4$ 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** $W_{Q_4,C_2}(E_1) \geq 3$ AND $CD_{Q_4,C_2}(E_1) = \text{“N”}$

**THEN** $adjW_{Q_4,C_2}(E_1) = W_{Q_4,C_2}(E_1) - 0.5$

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 $Q_3$ 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 $W_{Q_4,C_2}(E_1)$ is assigned to the value of $adjW_{Q_4,C_2}(E_1)$. After adjusting all weighting values for test item $Q_4$ and concept $C_2$, the set of those values of $n$ experts is defined as $adjW_{Q_4,C_2} = \{adjW_{Q_4,C_2}(E_i) | i = 1 \text{ to } n\}$.

**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, $adjW_{Q_4,C_2}$, calculate the density $dnstMX_{Q_4,C_2}$ of all values around the maximum value in the set as shown in Eq. (3):

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

(1)

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

(2)

$$dnstMX_{Q_4,C_2} = 1 - \frac{\max(adjW_{Q_4,C_2}) - \text{avg}(adjW_{Q_4,C_2})}{m}$$

(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 $dnstMN_{Q_4,C_2}$ of all values around the minimum as shown in Eq. (5):

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

(4)

$$dnstMN_{Q_4,C_2} = 1 - \frac{\text{avg}(adjW_{Q_4,C_2}) - \min(adjW_{Q_4,C_2})}{m}$$

(5)

The less the value of $dnstMX_{Q_4,C_2}$ compared to that of $dnstMN_{Q_4,C_2}$, the higher the probability that the majority of opinions from $n$ experts are closer to the value of $\min(adjW_{Q_4,C_2})$; that is, the value of $\max(adjW_{Q_4,C_2})$ is defined as a potential outlier, which represents an extreme weighting value among values for test item $Q_4$ and concept $C_2$ given by $n$ experts. The more the value of $dnstMN_{Q_4,C_2}$ compared to that of $dnstMN_{Q_4,C_2}$, the higher the probability that the majority of opinions from $n$ experts are closer to the value of $\min(adjW_{Q_4,C_2})$; that is, the value of $\max(adjW_{Q_4,C_2})$ is defined as a potential outlier.
for each sample size $n$. $\eta$ then equals half the range of the triangular distribution divided by this average MAD. That is, $\eta = \frac{\text{MAD}(\text{adjW}_{Q,C_4}) - \{\text{outlier}\}}{\text{MAD}}$ 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 $\text{adjW}_{Q,C_4}$. 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 $\eta$ when $n = 2$ to 15.

Step 4: After removing outliers, the integrated adjusted weighting value $iWeight_{Q,C_4}$ for test item $Q$ and concept $C_4$ is calculated as follows:

$$iWeight_{Q,C_4} = \text{avg}(\text{adjW}_{Q,C_4})$$

To verify that the value of $iWeight_{Q,C_4}$ is reasonable, the degree of majority opinion degree $dMO_{Q,C_4}$ needs to be calculated as follows:

$$dMO_{Q,C_4} = 1 - \frac{1}{m} \text{MAD}(\text{adjW}_{Q,C_4})$$

where $m$ represents the maximum rating scale (in this case, $m = 5$). The higher the value of $dMO_{Q,C_4}$, the higher probability that the experts have agreed on the value of $iWeight_{Q,C_4}$. A threshold, $\theta$, is used to indicate the acceptable agreement level. In this case, the value of $\theta$ is assumed to be 0.85. If $dMO_{Q,C_4} < \theta$, 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:

<table>
<thead>
<tr>
<th>$W_{Q,C_4}$</th>
<th>$CD_{Q,C_4}$</th>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
<th>$E_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>S</td>
<td>N</td>
<td>3</td>
<td>4</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
<td>N</td>
<td>N</td>
<td>3.5</td>
<td>4</td>
<td>S</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
<td>N</td>
<td>N</td>
<td>4</td>
<td>4</td>
<td>S</td>
</tr>
</tbody>
</table>

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:

$$\text{adjW}_{Q_2,C_4} = 2$$

Step 2: Based on the values of $\text{adjW}_{Q_2,C_4}$, given by, $\max \text{adjW}_{Q_2,C_4}$ is 4 and $\min \text{adjW}_{Q_2,C_4}$ is 2. Therefore the $\text{dnstMX}_{Q_2,C_4}$ and $\text{dnstMN}_{Q_2,C_4}$ values are evaluated as follows:

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

$$\text{dnstMX}_{Q_2,C_4} = 1 - \frac{4 - 3}{5} = 0.80$$

$$\text{dnstMN}_{Q_2,C_4} = 1 - \frac{3 - 2}{5} = 0.80$$

We can see that $\text{dnstMX}_{Q_2,C_4}$ is equal to $\text{dnstMN}_{Q_2,C_4}$; 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:

$$iWeight_{Q_2,C_4} = \text{avg}(2, 2.5, 3, 3.5, 4) = 3$$

To verify that the value of $iWeight_{Q_2,C_4}$ is reasonable, the value of $dMO_{Q_2,C_4}$ is calculated using Eq. (9) as follows:

$$dMO_{Q_2,C_4} = 1 - \frac{1}{5} \text{MAD}(2, 2.5, 3, 3.5, 4) = 0.88$$

We can see that the value of $dMO_{Q_2,C_4}$ is greater than 0.85; that is, the experts have agreed on the value $iWeight_{Q_2,C_4} = 3$, 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 $Q_3$ and concept $C_5$, 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:

<table>
<thead>
<tr>
<th>$W_{Q_3,C_5}$, $CD_{Q_3,C_5}$</th>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
<th>$E_5$</th>
<th>$E_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$, $S$</td>
<td>$0$</td>
<td>$0$</td>
<td>$N$</td>
<td>$1$, $S$</td>
<td>$1$, $S$</td>
<td>$2$, $S$</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>$adjW_{Q_3,C_5}$</th>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
<th>$E_5$</th>
<th>$E_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$</td>
<td>$0.5$</td>
<td>$1$</td>
<td>$1$</td>
<td>$2$</td>
<td>$2.5$</td>
<td></td>
</tr>
</tbody>
</table>

Step 2: Based on the values of $adjW_{Q_3,C_5}$ given by, max($adjW_{Q_3,C_5}$) is 2.5 and min($adjW_{Q_3,C_5}$) is 0. Therefore the $dnstMX_{Q_3,C_5}$ and $dnstMN_{Q_3,C_5}$ values are evaluated as follows:

\[
\text{avg}(\{0.5, 1, 1.2, 2.5\}) = \frac{7}{6}
\]

\[
dnstMX_{Q_3,C_5} = 1 - \frac{2.5 - \frac{7}{6}}{5} = 0.73
\]

\[
dnstMN_{Q_3,C_5} = 1 - \frac{\frac{7}{6} - 0}{5} = 0.77
\]

We can see that $dnstMN_{Q_3,C_5}$ is greater than $dnstMX_{Q_3,C_5}$; 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. $distOW_{Q_3,C_5}$ and $\text{MAD}(adjW_{Q_3,C_5} - \{\text{outlier}\})$ are calculated.

\[
distOW_{Q_3,C_5} = |2.5 - \text{avg}(\{0.5, 1, 1.2\})| = 1.60
\]

\[
\eta \times \text{MAD}(\{0.5, 1, 1.2\}) = 3.37 \times 0.52 = 1.75
\]

Because $distOW_{Q_3,C_5}$ is more than 1.25 ($1.60 > 1.25$), but it is still less than $\eta \times \text{MAD}(adjW_{Q_3,C_5} - \{\text{outlier}\})$ (1.60 < 1.75), the value 2.5 will be removed from the set $adjW_{Q_3,C_5}$. 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 $adjW_{Q_3,C_5}$ 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:

\[
iWeight_{Q_3,C_5} = \text{avg}(\{0.0, 0.5, 1, 1.2, 2.5\}) = 1.17
\]

To verify that the value of $iWeight_{Q_3,C_5}$ is reasonable, the value of $dMO_{Q_3,C_5}$ is calculated using Eq. (9) as follows:

\[
dMO_{Q_3,C_5} = 1 - \frac{1}{3} \text{MAD}(\{0.0, 0.5, 1, 1.2, 2.5\}) = 0.856
\]

We can see that the value of $dMO_{Q_3,C_5}$ 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 $C_3$ 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 $C_3$ given by seven experts are as follows:

<table>
<thead>
<tr>
<th>$W_{Q_1,C_3}$, $CD_{Q_1,C_3}$</th>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
<th>$E_5$</th>
<th>$E_6$</th>
<th>$E_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$, $S$</td>
<td>$1$, $N$</td>
<td>$4$, $N$</td>
<td>$4$, $S$</td>
<td>$5$, $N$</td>
<td>$5$, $N$</td>
<td>$5$, $S$</td>
<td></td>
</tr>
</tbody>
</table>

Step 1: Based on the elicited weighting values from seven experts, we can see that experts $E_2$, $E_5$, and $E_6$ have determined the weighting values for test item $Q_1$ and concept $C_3$ with low confidence. Therefore these weighting values are adjusted as follows:

<table>
<thead>
<tr>
<th>$adjW_{Q_1,C_3}$</th>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
<th>$E_5$</th>
<th>$E_6$</th>
<th>$E_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$</td>
<td>$1.5$</td>
<td>$3.5$</td>
<td>$4$</td>
<td>$4.5$</td>
<td>$4.5$</td>
<td>$5$</td>
<td></td>
</tr>
</tbody>
</table>

Step 2: Based on the values of $adjW_{Q_1,C_3}$ max($adjW_{Q_1,C_3}$) is 5 and min($adjW_{Q_1,C_3}$) is 0. Therefore the $dnstMX_{Q_1,C_3}$ and $dnstMN_{Q_1,C_3}$ values are evaluated as follows:

\[
\text{avg}(\{0.1, 0.5, 0.3, 0.4, 0.5, 0.5, 0.5\}) = 0.37
\]

\[
dnstMX_{Q_1,C_3} = 1 - \frac{5 - 3.7}{5} = 0.66
\]

\[
dnstMN_{Q_1,C_3} = 1 - \frac{23/7 - 0}{5} = 0.34
\]

We can see that $dnstMX_{Q_1,C_3}$ is greater than $dnstMN_{Q_1,C_3}$; 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. $\text{distOW}_{Q_i,c_5}$ and $\text{MAD}(\text{adjW}_{Q_i,c_5} - \{\text{outlier}\})$ are calculated.

$$\text{distOW}_{Q_i,c_5} = |0 - \text{avg}((1.5, 3.5, 4.4, 4.5, 5.5))| = 3.83$$
$$\eta \times \text{MAD}((1.5, 3.5, 4.4, 4.5, 5.5)) = 3.30 \times 0.89 = 2.93$$

Because $\text{distOW}_{Q_i,c_5}$ is more than 1.25 (3.83 > 1.25) and $\eta \times \text{MAD}(\text{adjW}_{Q_i,c_5} - \{\text{outlier}\})$ (3.83 > 2.93), the value 0 is removed from the set $\text{adjW}_{Q_i,c_5}$. Consequently, Step 2 is used once again to detect another outlier.

Step 2 (the 2nd round): After removing the outlier, the set $\text{adjW}_{Q_i,c_5} = \{1.5, 3.5, 4.4, 4.5, 5.5\}$. $\max(\text{adjW}_{Q_i,c_5})$ is 5 and $\min(\text{adjW}_{Q_i,c_5})$ is 1.5. Therefore the $\text{dnstMX}_{Q_i,c_5}$ and $\text{dnstMN}_{Q_i,c_5}$ values are evaluated as follows:

$$\text{avg}((1.5, 3.5, 4.4, 4.5, 5.5)) = \frac{23}{6}$$
$$\text{dnstMX}_{Q_i,c_5} = 1 - \frac{5 - 23/6}{5} = 0.77$$
$$\text{dnstMN}_{Q_i,c_5} = 1 - \frac{23/6 - 1.5}{5} = 0.53$$

The value of $\text{dnstMX}_{Q_i,c_5}$ is greater than $\text{dnstMN}_{Q_i,c_5}$; 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. $\text{distOW}_{Q_i,c_5}$ and $\text{MAD}(\text{adjW}_{Q_i,c_5} - \{\text{outlier}\})$ are calculated.

$$\text{distOW}_{Q_i,c_5} = |1.5 - \text{avg}((3.5, 4.4, 4.5, 4.5, 5.5))| = 2.80$$
$$\eta \times \text{MAD}((3.5, 4.4, 4.5, 4.5, 5.5)) = 3.37 \times 0.44 = 1.48$$

Because $\text{distOW}_{Q_i,c_5}$ is more than 1.25 (2.80 > 1.25) and $\eta \times \text{MAD}(\text{adjW}_{Q_i,c_5} - \{\text{outlier}\})$ (2.80 > 1.48), the outlier (1.5) is removed from the set $\text{adjW}_{Q_i,c_5}$. Consequently, Step 2 is used again to detect another outlier.

Step 2 (the 3rd round): After removing the outlier, the set $\text{adjW}_{Q_i,c_5} = \{3.5, 4.4, 4.5, 4.5, 5.5\}$. $\max(\text{adjW}_{Q_i,c_5})$ is 5 and $\min(\text{adjW}_{Q_i,c_5})$ is 3.5. Therefore the $\text{dnstMX}_{Q_i,c_5}$ and $\text{dnstMN}_{Q_i,c_5}$ values are evaluated as follows:

$$\text{avg}((3.5, 4.4, 4.5, 4.5, 5.5)) = 4.3$$
$$\text{dnstMX}_{Q_i,c_5} = 1 - \frac{5 - 4.3}{5} = 0.86$$
$$\text{dnstMN}_{Q_i,c_5} = 1 - \frac{4.3 - 3.5}{5} = 0.84$$

The value of $\text{dnstMX}_{Q_i,c_5}$ is greater than $\text{dnstMN}_{Q_i,c_5}$; 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. $\text{distOW}_{Q_i,c_5}$ and $\text{MAD}(\text{adjW}_{Q_i,c_5} - \{\text{outlier}\})$ are calculated.

$$\text{distOW}_{Q_i,c_5} = |3.5 - \text{avg}((4.4, 4.5, 4.5, 5.5))| = 1$$
$$\eta \times \text{MAD}((4.4, 4.5, 4.5, 5.5)) = 3.49 \times 0.25 = 0.87$$

Because $\text{distOW}_{Q_i,c_5}$ is less than 1.25 (1 < 1.25), the value 3.5 is not removed from the set $\text{adjW}_{Q_i,c_5}$. 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 $\text{adjW}_{Q_i,c_5}$ is $\{3.5, 4.4, 4.5, 4.5, 5.5\}$. Therefore, the integrated weighting value for test item $Q_i$ and concept $c_5$ is calculated by Eq. (8) as follows:

$$i\text{Weight}_{Q_i,c_5} = \text{avg}((3.5, 4.4, 4.5, 4.5, 5.5)) = 4.3$$

To verify that the value of $i\text{Weight}_{Q_i,c_5}$ is reasonable, the value of $d\text{MO}_{Q_i,c_5}$ is calculated using Eq. (9) as follows:

$$d\text{MO}_{Q_i,c_5} = 1 - \frac{1}{5} \text{MAD}((3.5, 4.4, 4.5, 4.5, 5.5)) = 0.91$$

We can see that the value of $d\text{MO}_{Q_i,c_5}$ 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 $Q_i$ 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 seven experts for test item $Q_i$ and concept $c_5$, 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 $Q_i$ and concept $c_5$ 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 $Q_i$ and concept $c_5$ given by eight experts are as follows:

<table>
<thead>
<tr>
<th>$W_{Q_i,c_5}$, $CD_{Q_i}$</th>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
<th>$E_5$</th>
<th>$E_6$</th>
<th>$E_7$</th>
<th>$E_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 0, 0, 0, 0, 0, 0, 0</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>N</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

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 $Q_i$ and concept $c_5$ with low confidence. Therefore these weighting values are adjusted as follows:

$$\text{adjW}_{Q_i,c_5} = \begin{cases} 0 & 0 \ 0.5 & 1 \ 4.5 & 5 \ 0.5 & 5 \ \end{cases}$$

Step 2: Based on the values of $\text{adjW}_{Q_i,c_5}$, given by, $\max(\text{adjW}_{Q_i,c_5})$ is 5 and $\min(\text{adjW}_{Q_i,c_5})$ is 0. Therefore the $\text{dnstMX}_{Q_i,c_5}$ and $\text{dnstMN}_{Q_i,c_5}$ values are evaluated as follows:

$$\text{avg}((0, 0, 0.5, 1, 4.5, 5.5, 5.5)) = 21/8$$
$$\text{dnstMX}_{Q_i,c_5} = 1 - \frac{5 - 21/8}{5} = 0.53$$
$$\text{dnstMN}_{Q_i,c_5} = 1 - \frac{21/8 - 0}{5} = 0.48$$

we can see that $\text{dnstMX}_{Q_i,c_5}$ is greater than $\text{dnstMN}_{Q_i,c_5}$; 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.
Step 4: set will be used for calculating an integrated weighting value in dMOQ test item less than g is not removed from the set value.

Step 3: Based on the value 0, verify whether it is distant enough from others and proper to be removed. distOWQ, and MAD(adjWQ) − (outlier) are calculated.

\[
\text{distOWQ} = |0 - \text{avg}([0, 0.5, 1, 4.5, 5, 5, 5])| = 3
\]

\[
\eta \times \text{MAD}([0, 0.5, 1, 4.5, 5, 5, 5]) = 3.26 \times 2.14 = 6.99
\]

Because \(\text{distOWQ}\) is more than 1.25 \((3 > 1.25)\), and but it is still less than \(\eta \times \text{MAD(adjWQ)} - \text{(outlier)}\) \((3 < 6.99)\), the outlier (0) is not removed from the set adjWQ. 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 adjWQ is \([0, 0, 0.5, 1, 4.5, 5, 5, 5]\). Therefore, the integrated weighting value for test item \(Q_0\) and concept \(C_7\) is calculated by Eq. (8) as follows:

\[
\text{iWeight}_{Q_0C_7} = \text{avg}([0, 0.0, 0.5, 1, 4.5, 5, 5, 5]) = 2.63
\]

To verify whether the value of iWeight_{Q_0C_7} is reasonable, the value of dMOQ_{Q_0C_7} is calculated using Eq. (9) as follows:

\[
dMOQ_{Q_0C_7} = 1 - \frac{1}{5} \text{MAD}([0, 0, 0.5, 1, 4.5, 5, 5, 5]) = 0.55
\]

We can see that the value of dMOQ_{Q_0C_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 \(Q_0\) 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

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Random weighting values with triangular distribution</th>
<th>Random confidence degree with uniform distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower limit (a)</td>
<td>Upper limit (b)</td>
</tr>
<tr>
<td>1</td>
<td>1 3 3 30 5 32.4 10.2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1 3 3 30 5 29.8 9.1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1 3 3 30 5 34.4 10.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 5
The parameters used to randomize dataset.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Data set</th>
<th># of expert</th>
<th># of item</th>
<th># of concept</th>
<th># of reconsiderations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Panjaburee et al. (2010) (%)</td>
</tr>
<tr>
<td>1</td>
<td>1 3 30 5 34.4</td>
<td>10.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 4 30 5 50.1</td>
<td>26.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 5 30 5 62.5</td>
<td>35.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 6 30 5 71.6</td>
<td>42.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 7 30 5 78.4</td>
<td>47.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 8 30 5 83.3</td>
<td>51.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7 9 30 5 87.0</td>
<td>54.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8 10 30 5 90.0</td>
<td>56.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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


