Journal of Educational Issues
ISSN 2377-2263
2016, Vol. 2, No. 2

A Model-Based Method for Content Validation of Automatically Generated Test Items

Xinxin Zhang (Corresponding author)

Department of Educational Psychology, University of Alberta

Edmonton T6G 2G5, Canada

Tel: 1-780-819-1773 E-mail: xinxin4@ualberta.ca

Mark Gierl

Department of Educational Psychology, University of Alberta

Edmonton T6G 2G5, Canada

Tel: 1-780-492-2396 E-mail: mark.gierl@ualberta.ca

Received: September 19, 2016 Accepted: October 26, 2016
Published: November 3, 2016

doi:10.5296/jei.v2i2.10040 URL: http://dx.doi.org/10.5296/jei.v2i2.10040

Abstract

The purpose of this study is to describe a methodology to recover the item model used to generate multiple-choice test items with a novel graph theory approach. Beginning with the generated test items and working backward to recover the original item model provides a model-based method for validating the content used to automatically generate test items. The methodology is demonstrated using items from a content area in medicine.

Keywords: item development; test development; automatic item generation; graph theory

1. Introduction

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) defines validity as: “the degree to which accumulated evidence and theory support specific interpretation of test scores entailed by proposed uses of a test.” This potentially large body of evidence can be categorized into content-related validity evidence, criterion-related validity evidence, and consequential-related validity evidence (Rudner & Schafer, 2002). Test content is one of the sources of validity evidence. The evidence is based on content relevance and representativeness of the items included in a test. This evidence is obtained from both judgmental and logical analyses of the test items, which is usually conducted by subject matter experts (SMEs) who are content specialists in the domain of interest. Validation is the process of evaluating the proposed interpretation of test scores (American Educational Research Association et al., 2014). Thus, content validation is a process used by SMEs to evaluate the relationship between a test's content and what the test is intended to measure by reviewing the items on a test in relation to their relevance for the domain of interest and the representativeness of the relevant items.

Automatic item generation (AIG) is an item development approach that uses both cognitive and psychometric theories to rapidly produce high-quality, content-specific test items, with the aid of computer technologies (Gierl & Haladyna, 2013). AIG relies on cognitive models designed by SMEs to produce new items through the use of algorithms that systematically organize and structure the item content. AIG is a three-step process consisting of cognitive model development, item model development, and item generation (Gierl & Lai, 2012). Typically, AIG content validation occurs after item generation. The purpose of content validation is to determine domain clarity, evaluate the items on a test in relation to their relevance and representativeness, and to make sure no errors have occurred in the presentation of the items during the generation process. It currently relies on a one-item-at-a-time review in which the SME evaluates the content of the generated items. Because AIG produces large numbers of items, SMEs usually sample the generated items for review. If low relevance, low representativeness, and/or presentation flaws are observed, then the SMEs will review another sample of generated items. The feedback will be collected for revising the item model. Then the newly generated items from the revised item model will be reviewed again to make sure the detected problems have been resolved in the generated items.

But the current approach to item review for content validation using generated items has two disadvantages. First, it is time consuming, especially when a large number of generated items are produced. Even though AIG is a breakthrough in the test development process because it satisfies the need for rapidly and efficiently producing large numbers of high-quality content-specific test items, its application of item review for content validation can still hinder the process. Suppose 2000 items, which is the minimum number of items for a 40-item computer adaptive test bank (cf. Breithaupt, Ariel, & Hare, 2009), are generated and ready for content validation. Estimated review time for each item by one SME, which includes collecting and recording the SME's judgments, is approximately 10 minutes. If three SMEs are involved in this process and they randomly review 70% of the items, then we can project that they would spend 42,000 minutes (700 hours) alone just to review one sample of the generated items. If flaws are detected during the process, then more items need to be viewed which will require even more time. Second, traditional item review has a high cost due to human capital. That is, the costs associated with the traditional item review method for content validation are severe. According to Statistics Canada, the average hourly wage for occupations in social science and education is $31.16. If we combine the above time estimation, then we can project that it would cost around $21,812 for content validation.

Furthermore, this projection is made under an unreal assumption that SMEs are fully satisfied with the generated items and offer no feedbacks or require no revisions. But if flaws are detected, then human capital costs increase.

Because of these important disadvantages, an alternative method which can save time and cost for content validation is needed. The purpose of the current research is to describe and illustrate an alternative method and to demonstrate this method with one practical and realistic example. The alternative method that will be described and demonstrated in the current study is validating AIG items through a recovery process that requires tracing the item model from the generated test items using graph theory. We call the new method an *AIG model-based review*. To-date, no one to our knowledge has used graph theory to validate item models for AIG in the review process of test development. Hence, the purpose of our study is to describe and illustrate this new method.

2. Literature Review

2.1 Item Development

Item development is one of the twelve essential, interrelated components required to create a test. The testing process starts with delineating an overall plan and concludes with producing test documentation to support its technical adequacy and validity (Lane, Raymond, Haladyna, & Downing, 2016). Item development involves activities like item writing, item content validation, item tryouts, and item banking, following the applicable standards to accumulate validity evidence to support and, sometimes, refute the intended interpretations and uses of test results.

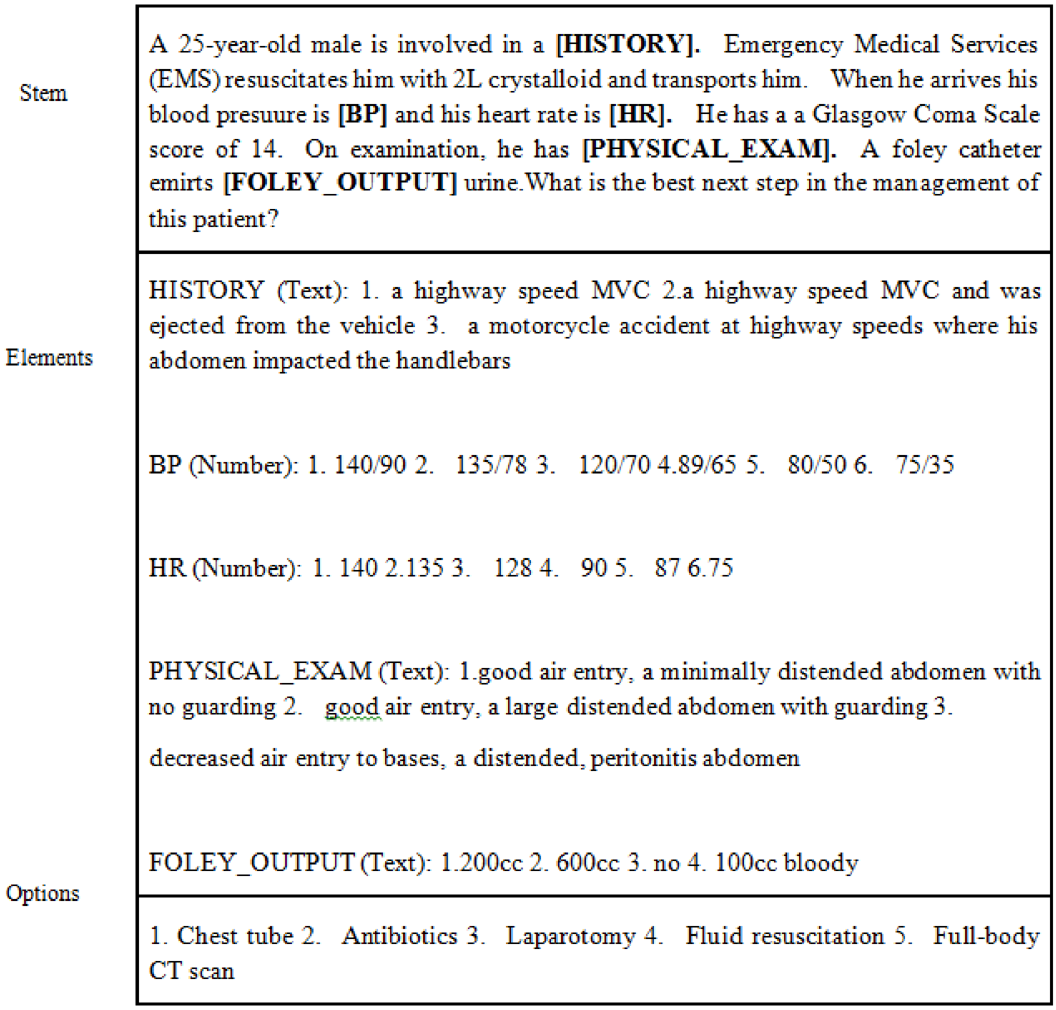
The traditional item development approach begins by recruiting and training SMEs to write items. They are responsible for locating related materials and creating items. Item writing is based on the judgement, experiences, and expertise of the SMEs. Once the items are developed, item content validation is conducted preferably by a new group of content experts who were not involved in developing the items. The reviewers evaluate the items on a test in relation to their relevance for the domain of interest and the representativeness of the relevant items. They also evaluate the printing, font size, and appropriateness of language. Depending on the outcomes from these reviews, some items are edited and reviewed again. Once items pass the content validation step, they are administered to a sample of examinees to evaluate the statistical properties. The items are typically evaluated for their difficulty and the extent to which they discriminate among examinees, which helps SMEs to decide which items will be retained for testing and which need to be deleted or revised. Item that meet high standards of quality are then securely stored in a database for use on operational exams.

2.2 Automatic Item Generation (AIG)

Automatic item generation (AIG) is a rapidly developing research area that offers an efficient way to generate items with the use of computer algorithms (Irvine & Kyllonen, 2002; Gierl & Lai, 2013). The role of the SMEs is not to locate materials and write individual items but to organize the resources and create meaningful item models for generating items. Gierl and Lai (2012) described a three-step process that includes developing cognitive model, creating item model, and generating items with the aid of computer technology. The first step in the AIG process is to develop cognitive models which highlight both the examinees' knowledge and skills required to solve the item as well as specify the content features in the items. To create the cognitive models, the SMEs are asked to identify and describe the key information that would be used to solve a test item (Gierl, Lai, & Turner, 2012). This cognitive model is used to guide the detailed rendering needed for item generation. The second step is to create item models. An item model is comparable to a template that contains the components in an assessment task, including the stem and the options. The specific variables in an item model are manipulated to produce new test items and the content used for these variables are identified in this step. Figure 1 shows an item model in the medical education domain. The upper box (stem-box) presents a stem with five variables [HISTORY], [BP], [HR], [PHYSICAL_EXAM] and [FOLEY_OUTPUT]. The middle box (element box) shows the corresponded content for these variables. The bottom box (option box) lists all the options including keys and distractors.

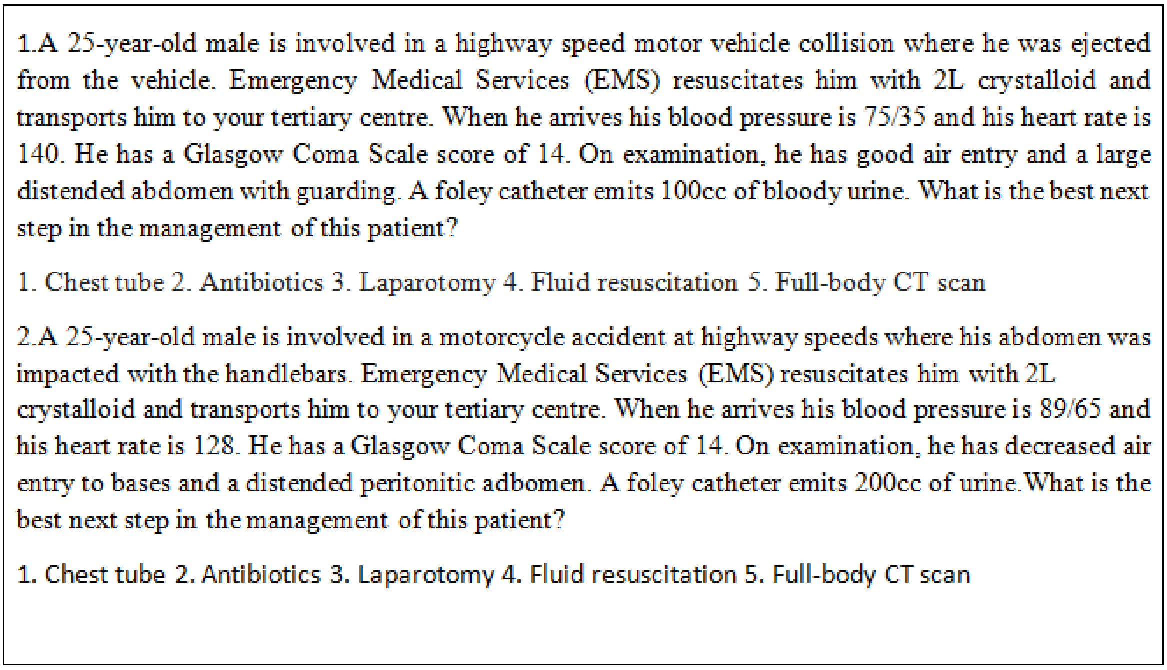
The third step is to generate items using computer software. All possible combinations of the variable content and options are assembled subject to the constraints articulated in the cognitive model, which ensures the generated items are sensible and useful. Two generated items from the item model above are presented in Figure 2.

Currently, generated items are validated by SMEs using the same process as with traditional item development. That is, item review relies on a one-item-at-a-time evaluation where SMEs scrutinize the content of the generated items. However, this item review approach is time consuming and costly, particularly when large numbers of new items must be reviewed. The AIG model-based review method presented in this research is designed to overcome these challenges.



Stem	A 25-year-old male is involved in a **[HISTORY].** Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him. When he arrives his blood presuure is **[BP]** and his heart rate is **[HR].** He has a a Glasgow Coma Scale score of 14. On examination, he has **[PHYSICAL_EXAM].** A foley catheter emirts **[FOLEY_OUTPUT]** urine.What is the best next step in the management of this patient?
Elements	HISTORY (Text): 1. a highway speed MVC 2.a highway speed MVC and was ejected from the vehicle 3. a motorcycle accident at highway speeds where his abdomen impacted the handlebars BP (Number): 1. 140/90 2. 135/78 3. 120/70 4.89/65 5. 80/50 6. 75/35 HR (Number): 1. 140 2.135 3. 128 4. 90 5. 87 6.75 PHYSICAL_EXAM (Text): 1.good air entry, a minimally distended abdomen with no guarding 2. good air entry, a large distended abdomen with guarding 3. decreased air entry to bases, a distended, peritonitis abdomen FOLEY_OUTPUT (Text): 1.200cc 2. 600cc 3. no 4. 100cc bloody
Options	1. Chest tube 2. Antibiotics 3. Laparotomy 4. Fluid resuscitation 5. Full-body CT scan

Figure 1. An item model with a stem and five options



1.A 25-year-old male is involved in a highway speed motor vehicle collision where he was ejected from the vehicle. Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary centre. When he arrives his blood pressure is 75/35 and his heart rate is 140. He has a Glasgow Coma Scale score of 14. On examination, he has good air entry and a large distended abdomen with guarding. A foley catheter emits 100cc of bloody urine. What is the best next step in the management of this patient?

1. Chest tube 2. Antibiotics 3. Laparotomy 4. Fluid resuscitation 5. Full-body CT scan

2.A 25-year-old male is involved in a motorcycle accident at highway speeds where his abdomen was impacted with the handlebars. Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary centre. When he arrives his blood pressure is 89/65 and his heart rate is 128. He has a Glasgow Coma Scale score of 14. On examination, he has decreased air entry to bases and a distended peritonitic adbomen. A foley catheter emits 200cc of urine.What is the best next step in the management of this patient?

1. Chest tube 2. Antibiotics 3. Laparotomy 4. Fluid resuscitation 5. Full-body CT scan

Figure 2. Two generated items using the item model in Figure 1

2.3 Graph Theory

The AIG model-based review is based on a graph theory analysis of the generated test items. Graph theory (GT) is the study of mathematical structures used to model pairwise relations between objects (Cavalcante, 2013). It is commonly used in mathematics and computer science (*e.g.*, Deo, 2004; Hammond, Vandergheynst, & Gribonval, 2011; Baker & Norine, 2007).

Graphs are used for graph theory research and analysis. Vertices (nodes) and edges (arcs) are the fundamental and indivisible units of a graph. A vertex *v* is expressed by a point or a circle. An edge is a link between two nodes. The edges may be directed or undirected. Directed edges connect ordered pairs of vertices where an arrow extending from one vertex to another vertex will be observed. Undirected edges connect unordered pairs of vertices by line. A directed graph is a graph whose edges are all directed. An undirected graph is a graph whose edges are all undirected. A graph with both directed and undirected edges is called a mixed graph (West, 2001). All three graph types are shown in Figure 3.

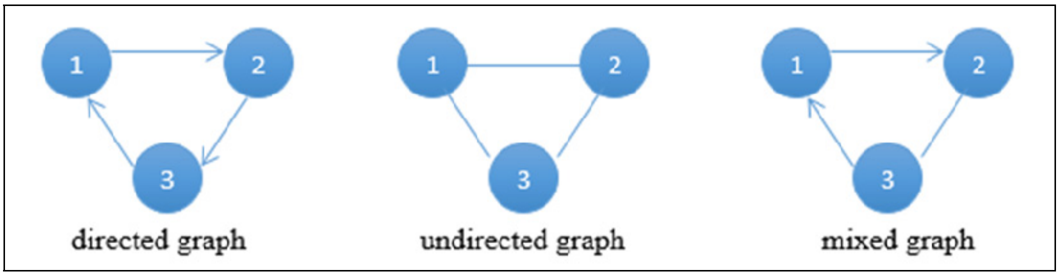


Figure 3. Examples of a directed graph, undirected graph, and mixes graph

A matrix is one of the data structures that can be used to represent a graph. Adjacency matrices specify the nodes' with adjacent relations. The adjacency matrix of a directed graph on *n* vertices is a *n* × *n* matrix where the diagonal entries a_{ij} are 0 and the non-diagonal entry a_{ij} can be 1, when there is an ordered edge from vertex *i* to vertex *j*. For a directed graph on 3 vertices, it has 3 × 3 adjacency matrix. The diagonal entries a_{11}, a_{22}, and a_{33}, are zeroes and the non-diagonal entries can be 0 or 1, depending on if there is a directed edge. Figure 4 shows a directed graph which has three vertices 1, 2, and 3 as well as three directed edges <1,2>, <2,3>, and <3,1>, and its 3 × 3 adjacency matrix, for which the entries of a_{12}, a_{23}, and a_{31} are 1 and the remaining entries are 0.

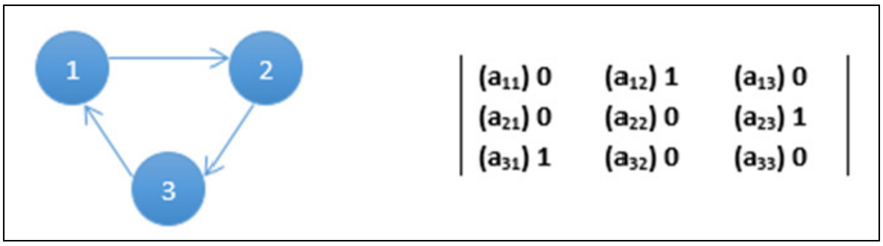
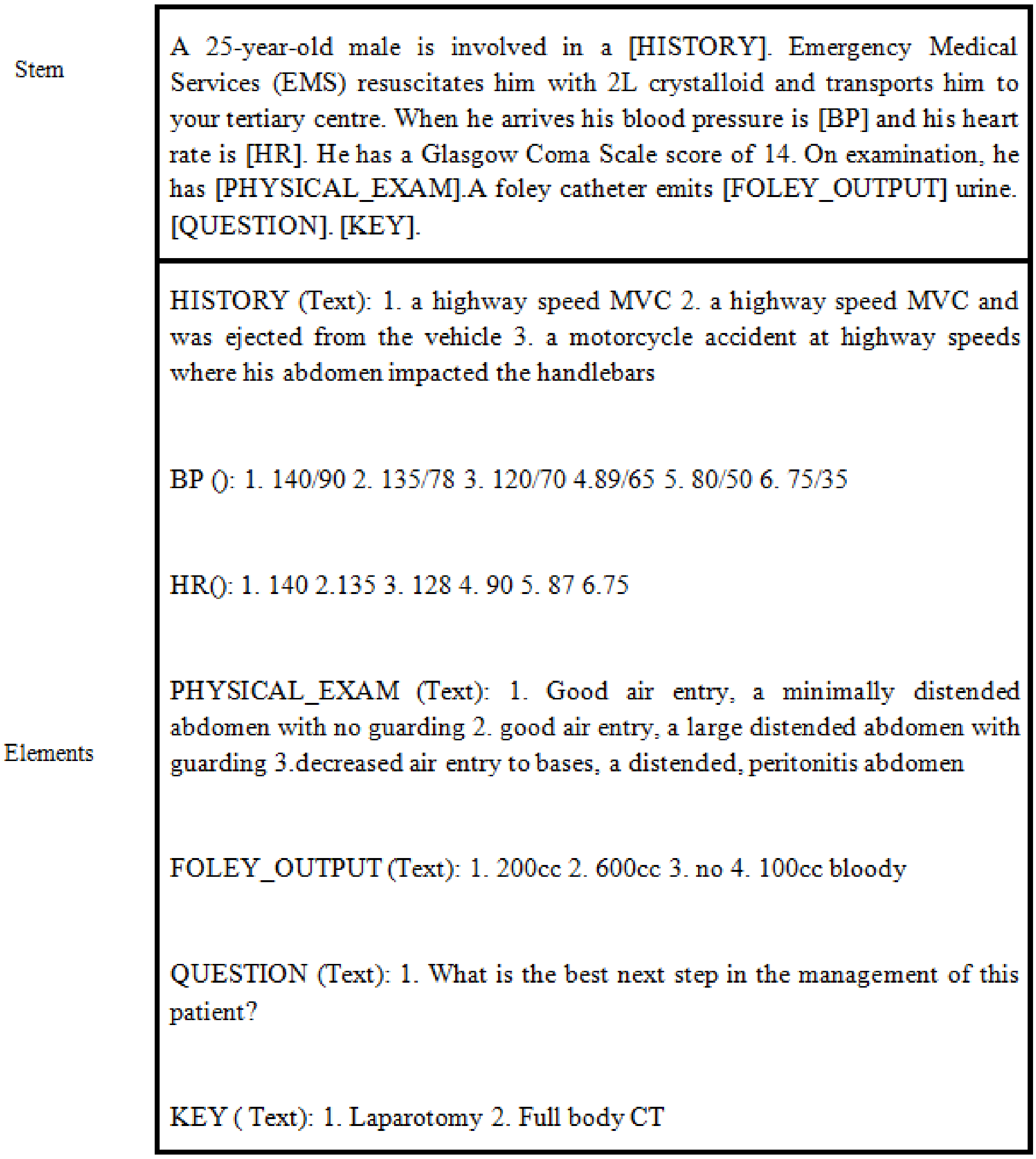


Figure 4. A directed graph and its adjacency matrix

2.4 Graph Theory and the AIG Model-Based Method

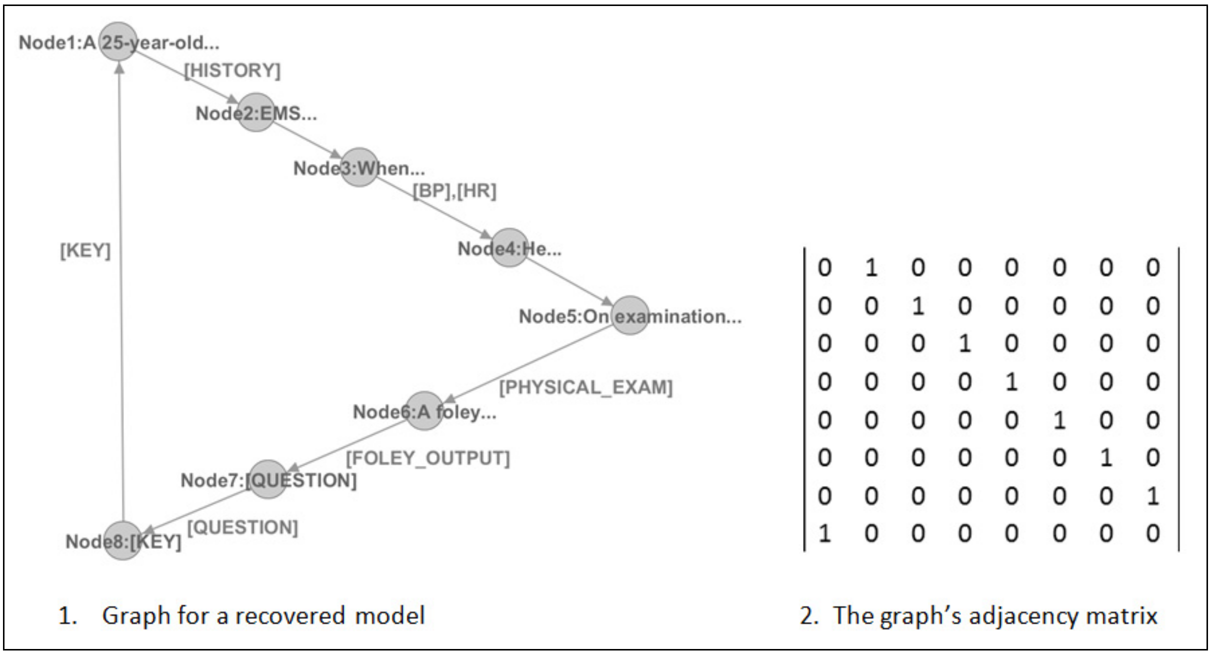
The graphical structure and its adjacency matrix are applied to present the recovered model from the generated items for review. The method for recovery will be discussed in more detail in the method section. In Figure 5, a recovered model is presented as an example. It is similar to the original item model, but with two more variables [QUESTION] and [KEY] in the stem-box. The corresponding values for these two variables are added in the element box and there is no option box.



Stem	A 25-year-old male is involved in a [HISTORY]. Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary centre. When he arrives his blood pressure is [BP] and his heart rate is [HR]. He has a Glasgow Coma Scale score of 14. On examination, he has [PHYSICAL_EXAM].A foley catheter emits [FOLEY_OUTPUT] urine. [QUESTION]. [KEY].
Elements	HISTORY (Text): 1. a highway speed MVC 2. a highway speed MVC and was ejected from the vehicle 3. a motorcycle accident at highway speeds where his abdomen impacted the handlebars BP (): 1. 140/90 2. 135/78 3. 120/70 4.89/65 5. 80/50 6. 75/35 HR(): 1. 140 2.135 3. 128 4. 90 5. 87 6.75 PHYSICAL_EXAM (Text): 1. Good air entry, a minimally distended abdomen with no guarding 2. good air entry, a large distended abdomen with guarding 3.decreased air entry to bases, a distended, peritonitis abdomen FOLEY_OUTPUT (Text): 1. 200cc 2. 600cc 3. no 4. 100cc bloody QUESTION (Text): 1. What is the best next step in the management of this patient? KEY (Text): 1. Laparotomy 2. Full body CT

Figure 5. Example of a recovered model

To present this model, we use a directed graph with eight nodes. Each of these nodes represents one sentence in the stem. As the first panel in Figure 6 shows, the first node is the first sentence of the stem “A 25-year-old male is involved in a [HISTORY]”. The second node is the second sentence “Emergency Medical Services (EMS) resuscitates him with 2 L crystalloid and transports him to your tertiary centre.” The third node is the third sentence “When he arrives his blood pressure is [BP] and his heart rate is [HR].” The fourth node is the fourth sentence “He has a Glasgow Coma Scale score of 14.” The fifth node is the fifth sentence “On examination, he has [PHYSICAL_EXAM].” The sixth sentence is the sixth node “A Foley catheter emits [FOLEY_OUTPUT] urine.” The seventh node is the seventh sentence “[QUESTION].” The eighth node is the eighth and last sentence “[KEY]”. The weight of each edge presents the variables contained in the edge’s initial node. For example, the weight of edge <Node1, Node2> is the variable [HISTORY], which is contained in Node1 “A 25-year-old male is involved in a [HISTORY]”. Panel 2 in Figure 6 is the adjacency matrix of this graph, which presents the nodes’ adjacent relations.



$$\begin{vmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{vmatrix}$$

1. Graph for a recovered model

2. The graph’s adjacency matrix

Figure 6. Examples of using graph to represent model

3. Method

An eight-step approach using graph theory was implemented to recover the model from the generated items. The generated items are multiple-choice items with a single stem and four options used to test medical students. The stem contains content (non-question component) and the test question. The options include a set of alternative answers with one correct option and three incorrect options. This methodology works with the stem and the correct option components thereby allowing the researcher to understand the relationship between the stem and the correct response using a multiple-choice item.

3.1 Step 1: Categorize the Items

The purpose of this step is to categorize the items based on the number of sentences. The items which have the same number of sentences are placed in the same category or “bin” which can be a sheet of an Excel file for further processing. The assumption is that the items with the same number of sentences might be generated from the same item model. Conversely, the items with a different number of sentences may be generated from a different item model. Based on this assumption, categorizing the items improves the efficiency of tracing the model.

3.2 Step 2: Parse the Items

The purpose of this step is to parse non-question component of the items in all Excel sheets. The Stanford Parser was used. The Stanford Parser is a program developed by the Natural Language Processing Group at Stanford University (2016). The parser identifies the grammatical structure of the sentences thereby allowing us to group words that go together as "phrases" and group words that are the subject or object of a verb. The example below presents a parsed sentence from a medical item. “A/DT 25-year-old/JJ male/NN is/VBZ involved/VBN in/IN a/DT highway/NN speed/NN MVC/NN.” Each word in the sentence is followed by a slash for separation and some capital letters which are the part-of-speech tags. Parts-of-speeches are the basic types of words in the English language that include nouns, verbs, adjectives, adverbs, pronouns, conjunctions, prepositions, and interjections. The Stanford Parser sets up its own part-of-speech tags for a single word according to its role in the sentence. In this example, DT stands for determiner, JJ stands for adjective, NN stands for noun, VBZ stands for verb’s 3rd single singular present, VBN stands for verb’s past tense, and IN stands for preposition. The outcome of this step is parsed sentences with identified grammatical structures.

3.3 Step 3: Restate the Items

The purpose of this step is to restate each parsed sentence based on the grammatical structure. A grammatical link was used to realize this purpose. The grammatical link consists of part-of-speech tags and space. Take this parsed sentence as an example: “A/DT 25-year-old/JJ male/NN is/VBZ involved/VBN in/IN a/DT highway/NN speed/NN MVC/NN.” The basic grammatical structure of this sentence is: subjective-verb “male/NN is/ VBZ involved/VBN”. “Male” is the subjective, and “is involved” is the verb. The phrases “A 25-year-old” and “in a highway speed MVC” separately modifies the subjective “male” and the verb “is involved”. The grammatical link keeps the part-of-speech tags of the basic grammatical structure “male/NN is/ VBZ involved/VBN” and uses the space to replace the modification components “a 25-year-old” and “in a highway speed MVC”. Thus the grammatical link of this sentence becomes “() NN VBZ VBN ()”. There are two reasons for using a grammatical link. First, it improves the efficiency of recovering the item model. The sentences which have the same grammatical link are more likely generated from the same item model. Second, the grammatical link is programming friendly because it flags the locations for matching. The importance of this point will be discussed later in the methods section. The outcome of this step is a grammatical link for each parsed sentence.

3.4 Step 4: Get the Abstracted Pattern

The purpose of this step is to get the abstracted pattern for each sentence. The abstracted pattern highlights what the sentence looks like and where the specific variables are located in the sentence. To produce the outcome in this step, the sentences with the same grammatical link are gathered together for tracing the abstracted pattern through matching. To demonstrate the logic of this step, two sentences with the same grammatical link are used to illustrate this concept. The two sentences are: “A 25-year-old male is involved in a highway speed MVC.” and “A 25-year-old male is involved in a motorcycle accident at highway speeds where his abdomen impacted the handlebars.” Their common grammatical link is “() NN VBZ VBN ()”. The space of this grammatical link identifies and isolates where to compare and where to match. In the current example, the first space directs a comparison between the modification components “a 25-year-old” and “a 25-year-old”. The second space directs a comparison between the modification components “in a highway speed MVC” and “in a motorcycle accident at highway speeds where his abdomen impacted the handlebars”. Then the corresponding words from the part-of speech tag in two sentences are compared separately. The first part-of-speech “NN” directs a comparison between “male” and “male”. The second part-of speech tag “VBZ VBN” directs a comparison between “is involved” and “is involved”. By keeping the same words and replacing the different word/phrases with brackets, an abstracted pattern “A 25-year-old male is involved in [ ]” is produced for these two sentences. The bracket indicates a variable. Then, the different phrases “a highway speed MVC” and “a motorcycle accident at highway speeds where his abdomen impacted the handlebars” are recorded for further processing. The outcome of this step is the abstracted pattern for each sentence.

3.5 Step 5: Develop the Structure Table

The purpose of this step is to develop the structure table. In order to develop this table, two sub-steps are required. First, the abstracted patterns are combined and listed in this table. Second, two abstracted patterns-“[Question]” and “[Key]” are added together to create a list that includes all of the information in a test item. Table 1illustrates an outcome of this step. This structure table has nine abstracted patterns. The eighth is for question and the ninth is for the key.

Table 1. A structure table

No	Abstracted Pattern
1	A 25-year-old male is involved in a [1].
2	Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary center.
3	When he arrives his blood pressure is [2] and his heart rate is [3].
4	He has a Glasgow Coma Scale score of 14.
5	He is complaining of lower-rib pain on his [4].
6	On examination, he has [5].
7	A Foley catheter emits [6] urine.
8	[QUESTION].
9	[KEY].

3.6 Step 6: Develop the Content Table

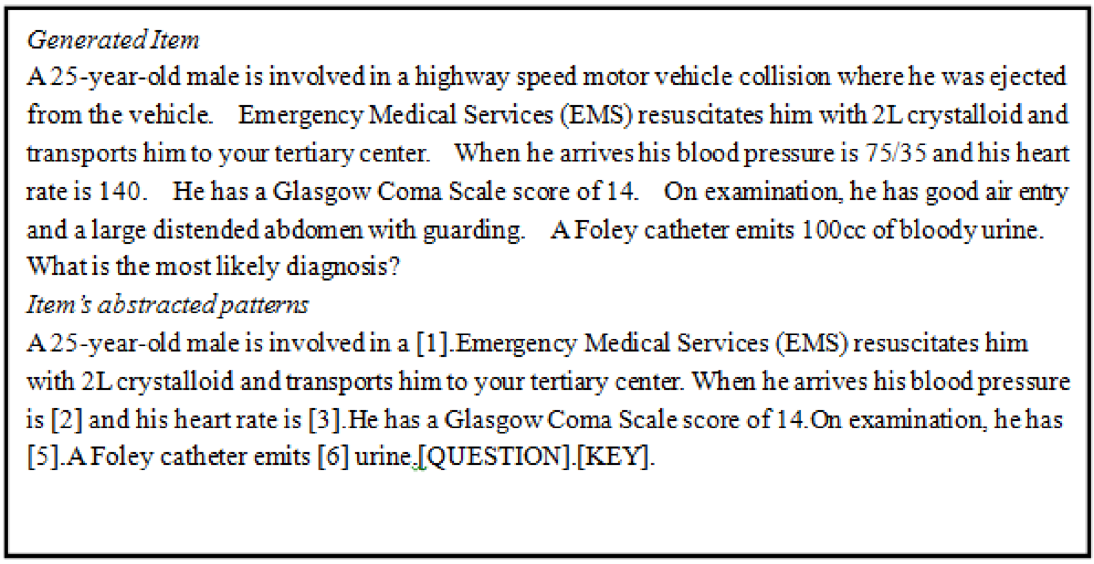
The purpose of this step is to develop the content table, which lists the content for the variables in the structure table in step 5. The recorded word/phrases from step 4 are listed in this table as the content. Continuing with the previous example, for the first abstracted pattern in the structure table "A 25-year-old male is involved in [1]." the recorded phrases "a highway speed MVC" and "a motorcycle accident at highway speeds where his abdomen impacted the handlebars" in step 4 are listed in the content table as variable [1]'s content. Table 2 illustrates one sample outcome of this step—a content table using the recorded information for each variable that is presented in Table1.

Table 2. A content table

Variable	Conent
[1]	1. a highway speed MVC 2. a highway speed MVC and was ejected from the vehicle 3. a motorcycle accident at highway speeds where his abdomen impacted the handlebars
[2]	1. 140/90 2. 135/78 3. 120/70 4.89/65 5. 80/50 6. 75/35
[3]	1. 140 2.135 3. 128 4. 90 5. 87 6.75
[4]	1. Good air entry, a minimally distended abdomen with no guarding 2. good air entry, a large distended abdomen with guarding 3.decreased air entry to bases, a distended, peritonitis abdomen
[5]	1. right side 2. left side
[6]	1. 200cc 2. 600cc 3. no 4. 100cc bloody
[Question]	1. What is the best next step in the management of this patient? 2.What is the most likely diagnosis?
[Key]	1. Laparotomy 2. Full body CT 3. Splenic rupture

3.7 Step 7: Generate Sequences

The purpose of this step is to list the structure for the items using sequences. This step is used to identify the structure sequence by matching the structure table in step 5 to the items' abstracted patterns in step 4. A generated item with its abstracted patterns is given in Figure 7 to demonstrate this concept. The outcome of this step is a sequence "1.2.3.4.6.7.8.9". This sequence represents the item's structure. Each number in this sequence corresponds to an abstracted pattern listed in the structure table in Table1. For example, the 5th number "6" corresponds to the 6th abstracted pattern "On examination, he has [ ]." of Table 1. The outcome of this step is the structure sequences, which describe the recovered item model.



Generated Item

A 25-year-old male is involved in a highway speed motor vehicle collision where he was ejected from the vehicle. Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary center. When he arrives his blood pressure is 75/35 and his heart rate is 140. He has a Glasgow Coma Scale score of 14. On examination, he has good air entry and a large distended abdomen with guarding. A Foley catheter emits 100cc of bloody urine. What is the most likely diagnosis?

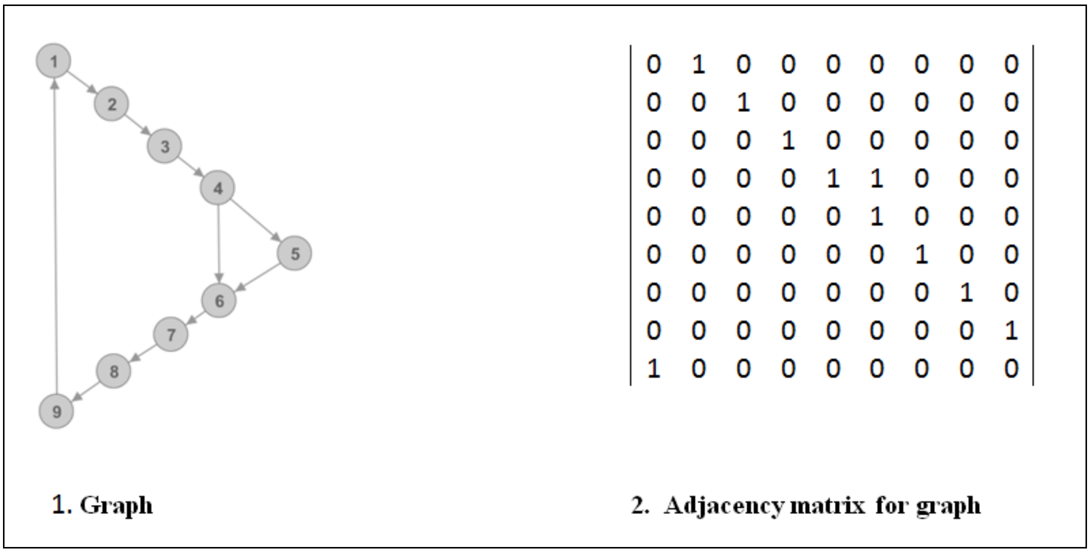
Item's abstracted patterns

A 25-year-old male is involved in a [1].Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary center. When he arrives his blood pressure is [2] and his heart rate is [3].He has a Glasgow Coma Scale score of 14.On examination, he has [5].A Foley catheter emits [6] urine.[QUESTION].[KEY].

Figure 7. A generated item and its abstracted patterns

3.8 Step 8: Apply Graph Theory

The purpose of this step is to use graph theory to create the recovered model. Two sub-steps are taken. First, the graph is used to describe the recovered model. Then, the adjacency matrix is used to describe the graph. In sub-step 1, the nodes of the graph are used to express the structure sequence developed from step 7. Panel 1 in Figure 8 presents a graph with nine nodes. The two paths of this graph present two structure sequences “1.2.3.4.6.7.8.9” and “1.2.3.4.5.6.7.8.9“. This graph indicates the structure for all generated items. In other words, all the generated items are from the model with two paths “1.2.3.4.6.7.8.9” and “1.2.3.4.5.6.7.8.9”. In sub-step 2, the graph’s adjacency matrix is produced in panel 2. The outcome of this sub-step is the graph and adjacency matrix.



$$\left|\begin{matrix}0&1&0&0&0&0&0&0&0\\0&0&1&0&0&0&0&0&0\\0&0&0&1&0&0&0&0&0\\0&0&0&0&1&1&0&0&0\\0&0&0&0&0&1&0&0&0\\0&0&0&0&0&0&1&0&0\\0&0&0&0&0&0&0&1&0\\0&0&0&0&0&0&0&0&1\\1&0&0&0&0&0&0&0&0\end{matrix}\right|$$

1. Graph

2. Adjacency matrix for graph

Figure 8. An example of the outcome from step 8

4. Results

Three different sets of items from the medical education domain were generated and the model was recovered to demonstrate our method. Due to space limitation, we only present the first example in the Results section. The outcomes from the second and third examples can be obtained from the first author. All the generated items were created by medical SMEs using the three-step AIG process that included developing a cognitive model, creating an item model, and generating items (Gierl, Lai, &Turner, 2012). The items in each dataset were generated based on individual cognitive models and the derived item models. The first dataset has 938 items related to abdominal trauma.

After applying the 8-step methodology for recovery, the results for the model structure table, content table, graph, and graph matrix were produced and are presented in this section. The structure table identifies what components (*i.e.*, content, question, and key) are in the model, how the components are structured, and where the specific variables are located. The content table further specifies the content of the specific variables. The graph is a visual representation of the item model. The graph matrix is a summary of the graphical relations.

Table 3 is the structure table developed in step 5 of recovering the model in the abdominal trauma dataset. It lists nine abstracted patterns separately in rows. This table identifies the three components in the model. The first to seventh abstracted patterns (row 1 to row 7) are the content (non-question) component. The eighth abstracted pattern (row 8) is the question component and the ninth abstracted pattern (row 9) is the key component. The brackets indicate the variables in the model. The model in the abdominal trauma dataset has eight variables in total. Variable [1] to [6] are located in the content component. They include variable [1] presented in the first abstracted pattern, variable [2] and [3] presented in the third abstracted pattern, variable [4] presented in the fifth abstracted pattern, variable [5] presented in the sixth abstracted pattern, and variable [6] presented in the seventh abstracted pattern. The seventh variable [question] is located in the question component (the eighth abstracted pattern) and the eighth variable [key] is located in the key component (the ninth abstracted pattern).

Table 3. Structure table for abdominal trauma dataset

No	Abstracted Pattern
1	A 25-year-old male is involved in a [1].
2	Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary center.
3	When he arrives his blood pressure is [2] and his heart rate is [3].
4	He has a Glasgow Coma Scale score of 14.
5	He is complaining of lower-rib pain on his [4]
6	On examination, he has [5].
7	A Foley catheter emits [6] urine.
8	[QUESTION].
9	[KEY].

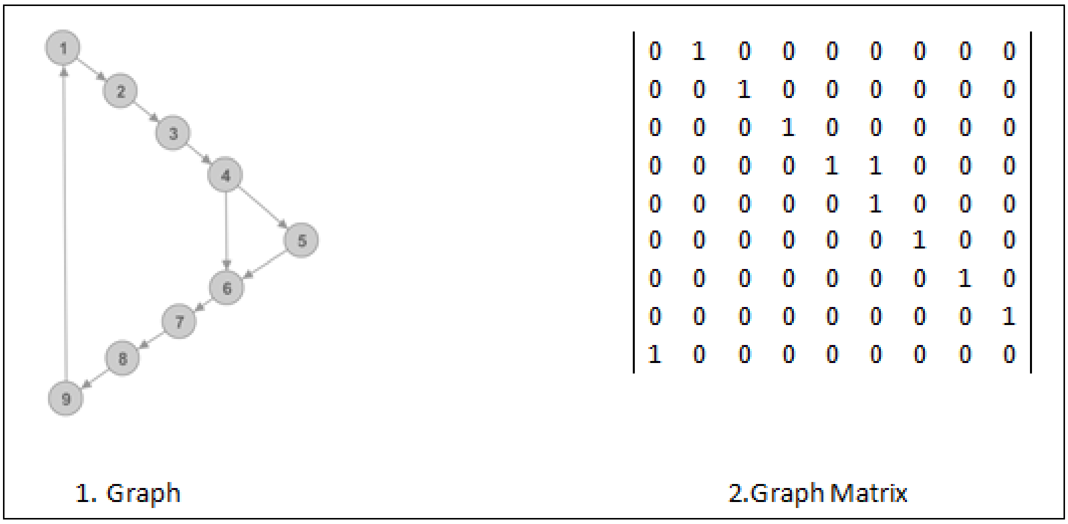
Table 4 is the content table developed in the step 6. It lists the content for the variables in Table 3. For example, variable [2] in the third abstracted pattern of Table 3 has six values varying from “1. 140/90” to “6.75/35”. Variable [key] in the ninth abstracted pattern of Table 3 has three values varying from “1. Laparotomy” to “3. Splenic rupture”.

Table 4. Content table for abdominal trauma dataset

Variable	Conent
[1]	1. a highway speed MVC 2. a highway speed MVC and was ejected from the vehicle 3. a motorcycle accident at highway speeds where his abdomen impacted the handlebars
[2]	1. 140/90 2. 135/78 3. 120/70 4.89/65 5. 80/50 6.75/35
[3]	1. 140 2.135 3. 128 4. 90 5. 87 6. 75
[4]	1. Good air entry, a minimally distended abdomen with no guarding 2. good air entry, a large distended abdomen with guarding 3.decreased air entry to bases, a distended, peritonitis abdomen
[5]	1. right side 2 left side
[6]	1. 200cc 2. 600cc 3. no 4. 100cc bloody
[QUESTION]	1. What is the best next step in the management of this patient? 2.What is the most likely diagnosis?
[KEY]	1. Laparotomy 2. Full body CT 3. Splenic rupture

Panel 1 in Figure 9 is the graph developed in the step 8. It structures the generated items from the abdominal trauma dataset. In other words, it presents the recovered model in abdominal trauma dataset. This graph has two paths which are "1.2.3.4.6.7.8.9" and "1.2.3.4.5.6.7.8.9". It indicates that all the generated items in the abdominal trauma dataset are from the model with the structure sequences "1.2.3.4.6.7.8.9" and "1.2.3.4.5.6.7.8.9". Based on Table 3, we know the model has two paths. The first path is "A 25-year-old male is involved in a [1]. Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary center. When he arrives his blood pressure is [2] and his heart rate is [3]. He has a Glasgow Coma Scale score of 14. On examination, he has [5]. A Foley catheter emits [6] urine. [QUESTION].[KEY]." The second path is "A 25-year-old male is involved in a [1]. Emergency Medical Services (EMS) resuscitates him with 2L crystalloid and transports him to your tertiary center. When he arrives his blood pressure is [2] and his heart rate is [3]. He has a Glasgow Coma Scale score of 14. He is complaining of lower-rib pain on his [4]. On examination, he has [5]. A Foley catheter emits [6] urine. [QUESTION].[KEY]".

Panel 2 in Figure 9 is the adjacency matrix for the graph in panel 1, developed in step 8 of the recovering process. This 9*9 matrix specifies the adjacent relations of the nine vertexes in the graph. The non-diagonal entries a_{12},a_{23}, a_{34}, a_{45}, a_{46}, a_{56}, a_{67}, a_{78},a_{89},a_{91} with 1 indicate 10 ordered edges from vertex 1 to vertex 2, vertex 2 to vertex 3, vertex 3 to vertex 4, vertex 4 to vertex 5, vertex 4 to vertex 6, vertex 5 to vertex 6, vertex 6 to vertex 7, vertex 7 to vertex 8, vertex 8 to vertex 9, and vertex 9 to vertex 1.



$$\begin{vmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{vmatrix}$$

Figure 9. A graph and a graph matrix for abdominal trauma dataset

5. Conclusion and Discussion

Automatic item generation is a new approach for item development that satisfies a testing agencies' requirement to produce large numbers of high-quality items in a timely and cost-effective manner. With the aid of the computer, cognitive and item models are used to generate items. After the items are generated, the one-item-at-a-time validation method is used by SMEs to review the generated items in order to analyze the relationship between the content and what the item is intended to measure. But this validation method is time consuming and costly, particularly when large numbers of new items must be reviewed. In order to overcome these challenges, a model-based validation method was developed and demonstrated in this study.

Using the proposed method, large numbers of generated items can be validated by reviewing the structure table, content table, graph, adjacency matrix or/and variable content sequences. These outcomes provide the SME with important benefits during the review process. The structure table lists all the abstracted patterns which are used to evaluate the main concept, its associated scenarios, and the information resources within each abstracted pattern. The content table specifies the content of the variables in the structure tables which are used to evaluate the appropriateness of the content and the accuracy of the presentation. The graph structures the item is used to evaluate the individual task structure. Depending on practical needs, SMEs can review any combinations of the products for validation. For example, if they focus more on the structure of the dataset than the concrete content, then they can just review the graph. After the model-based validation, feedback will be provided to the original AIG model developer for improving their item model. This approach allows the SME to evaluate the information used to generate the items rather than focusing on the items themselves. Hence, all of the content use for item generation can be scrutinized and evaluated rather than relying on a review of samples of generated items.

The model-based validation method is a recovery process that starts with generated items and ends with model, however the cognitive and item modeling step in AIG is a development process that starts with model and ends with items. Cognitive modeling requires the development of a structure that specifies the knowledge and skills required to solve test items which leads to the creation of new items. By comparison, the validation method begins with the generated test items and works backward to recover the original item model using a systematic process supported by graph theory analysis. The model-based validation method is a solution to the challenging problem of item review when large numbers of generated items are created. Using this method, the SMEs can avoid reviewing the content of every selected item. Instead, they review the summarized products extracted from the items, which saves both time and effort, in the process of producing large numbers of high-quality test items.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Baker, M., & Norine, S. (2007). Riemann–Roch and Abel–Jacobi theory on a finite graph. *Advances in Mathematics, 215*(2), 766-788. http://dx.doi.org/10.1016/j.aim.2007.04.012

Breithaupt, K., Ariel, A. A., & Hare, D. R. (2009). Assembling an inventory of multistage adaptive testing systems. *Elements of adaptive testing* (pp. 247-266). Springer New York. http://dx.doi.org/10.1007/978-0-387-85461-8_13

Cavalcante, A. (2013). *Graph Theory: New Research*. Hauppauge, New York: Nova Science Publishers, Inc.

Deo, N. (2004). *Graph theory with applications to engineering and computer science*. PHI Learning Pvt. Ltd., India.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing, 12*(3), 273-298. http://dx.doi.org/10.1080/15305058.2011.635830

Gierl, M. J., & Lai, H. (2013). Instructional Topics in Educational Measurement (ITEMS) Module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32*(3), 36-50. http://dx.doi.org/10.1111/emip.12018

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757-765. http://dx.doi.org/10.1111/j.1365-2923.2012.04289.x

Hammond, D. K., Vandergheynst, P., & Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis, 30*(2), 129-150. http://dx.doi.org/10.1016/j.acha.2010.04.005

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Hillsdale, NJ: Erlbaum. http://dx.doi.org/10.1016/j.acha.2010.04.005

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2015). *Handbook of test development*. Routledge.

Rudner, L. M., & Schafer, W. D. (2002). *What teachers need to know about assessment*. Washington, DC: National Education Association.

The Stanford Natural Language Processing Group. (n.d.). *The Stanford Parser: A statistical parser*. Retrieved March 30, 2016, from http://nlp.stanford.edu/software/lex-parser.shtml

West, D. B. (2001). *Introduction to graph theory* (Vol. 2). Upper Saddle River: Prentice hall.

Copyright Disclaimer