Education Tech Research Dev (2011) 59:741–763
DOI 10.1007/s11423-010-9183-0

RESEARCH ARTICLE

Measuring learning in serious games: a case study with structural assessment

Pieter Wouters · Erik D. van der Spek · Herre van Oostendorp

Published online: 27 February 2011

Abstract The effectiveness of serious games is often measured with verbal assessment. As an alternative we propose Pathfinder structural assessment (defined as measuring the learners' knowledge organization and compare this with a referent structure) which comprises three steps: knowledge elicitation, knowledge representation and knowledge evaluation. We discuss practical and theoretical considerations for the use of structural assessment and showcase its application with the game Code Red: Triage. Results suggest that structural assessment measures an individual's understanding of a domain at least differently from verbal assessment. While verbal assessment may provide a more nuanced picture regarding declarative and procedural knowledge, structural assessment may add an in-depth understanding of the concepts that are regarded important in a domain. In the Discussion we propose four guidelines to effectively use structural assessment in serious games: (1) Determine the appropriateness of the domain for structural assessment, (2) select an appropriate referent for the target group(s), (3) select the number of concepts needed for structural assessment, and (4) consider the analysis of the graphical knowledge representations to obtain in-depth information about the quality of the knowledge structures.

Keywords Serious games · Complex skills · Knowledge structures · Structural assessment · Verbal assessment

Introduction

Traditionally, verbal assessment (e.g., knowledge tests, transfer tests) is used in educational research to determine the effectiveness of a learning environment.

P. Wouters (✉) · E. D. van der Spek · H. van Oostendorp
Institute of Information and Computing Sciences, Utrecht University,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
e-mail: pieterw@cs.uu.nl

Nowadays we see an increasing use of computer games in learning and instruction (henceforth referred to as serious games)—often to train complex skills in realistic contexts- in which verbal assessment is also predominant (Wouters et al. 2009). Learning complex skills starts with the acquisition of declarative knowledge, but with increasing expertise, the accumulation of declarative knowledge is not sufficient and the focus will gradually shift to the construction of procedural knowledge and the organization of knowledge in meaningful structures (Kraiger et al. 1993). Current thinking in cognitive science emphasizes the importance of knowledge organization in learning complex skills as a counterpart for the amount of (declarative) knowledge that is learned (Day et al. 2001; Dorsey et al. 1999; Kraiger et al. 1993). An adequate organization of knowledge warrants the capability to integrate new information with existing knowledge making the new information meaningful. In other words, effective knowledge structures provide a context in which objects, information and events during training can be interpreted (Messick 1984). Secondly, research regarding the novice-expert shift has shown that an effective knowledge structure will enable faster problem solving because it not only contains the concepts (i.e., nodes) required to identify the characteristics of a problem, but also the concepts that can be used to solve the problem (Glaser and Chi 1989). We propose that structural assessment can be used to measure the quality of knowledge structures.

The underlying assumption of the structural assessment approach is that we organize our knowledge in knowledge structures containing the important concepts of a domain and the relations among those concepts (Kraiger et al. 1993). Changes in the knowledge structures due to learning or training have been investigated with techniques such as the ordered-tree techniques (Jonassen et al. 1993), hierarchical cluster analysis (Adelson 1981), relationship-judgment tests (Diekhoff 1983), concept maps (Keppens and Hay 2008), multidimensional scaling (Gonzalvo et al. 1994) and network techniques (Goldsmith et al. 1991). Also efforts have been made to integrate and fully automate the different steps in the assessment of knowledge structures. Examples are HiMatt (see Pirnay-Dummer et al. 2010) and Pathfinder (see Schvaneveldt et al. 1985). In this study we adopt Pathfinder for four reasons. First, unlike techniques like cluster analysis and ordered-tree technique it does not force a hierarchical structure on the data but identifies meaningful links between concepts (Jonassen et al. 1993). Second, we are interested in local (pairwise) relations between concepts rather than more global or dimensional relations (such as in multidimensional scaling). Third, it is a rather straightforward method with predefined concepts, requiring less introspection and cognitive effort from the assessee (compared for example with cognitive maps, cf. Trumpower et al. 2010). Fourth, research has shown that the Pathfinder technique is a valid and reliable procedure (cf. Dorsey et al. 1999; Gonzalvo et al. 1994).

In this study we explore if structural assessment with Pathfinder can be used to assess learning complex skills with serious games. For this purpose we also compare it with traditional verbal assessment in order to see whether structural assessment indeed reveals additional information. Structural assessment with Pathfinder is rather unknown in educational research; therefore we first describe the underlying theory, the required steps to perform a Pathfinder structural assessment and some theoretical and practical considerations. Then we showcase the application of structural assessment on learning with the game 'Code Red: Triage' (van der Spek et al., in press). Finally, we formulate potential avenues for future research and provide some tentative guidelines for the use of structural assessment with Pathfinder.

Structural assessment

Three steps can be distinguished to implement a structural approach to measure and interpret knowledge structures.

(1) Knowledge elicitation

The first step involves the *elicitation of knowledge*. This can be realized by a variety of methods such as card sorting, ordered recall or numerically rating the degree of relatedness between concepts (Goldsmith et al. 1991). Rating the relation between pairs of concepts is often applied in cognitive psychology and is assumed to capture the underlying organization of knowledge (Gonzalvo et al. 1994). For this purpose an analysis has to be made of important concepts in the domain or task, for example by conducting a cognitive task analysis, analyzing instructional material, interviewing experts/instructors or simply considering the agreed core concepts in a domain (Trumpower et al. 2010). However, before selecting the important concepts the question should be raised which domains are suitable for structural assessment and which are not. From the literature we have deduced two considerations. First, structural assessment is derived from proximities for pairs of entities. Proximity is the term used to refer to a measure of relationship between entities. This measure can be relatedness, distance or association. Although entities are usually concepts, they can be anything with a pattern of relationship (Schvaneveldt et al. 1985, p. 303). The prevailing view on knowledge structures is that they are associative networks of ideas, concepts, procedures and other forms of knowledge (cf. Trumpower et al. 2010). This view implies that structural assessment can also be used for less conceptually-oriented domains such as domains with a focus on procedures. Second, it is important that experts in a domain agree about the essential concepts and relations in a domain. Agreement is more likely to occur in domains with a central body of theory which is generally agreed upon (i.e., 'hard' domains such as physics) than 'soft' domains (e.g., history) that lack such central body of theory (Biglan 1973; Keppens and Hay 2008).

It seems desirable to create a large sample of concepts in order to have a comprehensive representation of an individual's knowledge structure. However, with n concepts participants have to conduct $n(n-1)/2$ pair wise ratings which can easily lead to an unrealistic number of ratings. Studies have used concept sample sizes ranging from 6 (Diekhoff 1983) to 32 (Gonzalvo et al. 1994). Although Goldsmith et al. (1991) found that the higher the number of concepts, the better the performance on an exam was predicted, we agree with Trumpower et al. (2010) that too many concepts may compromise the validity of structural assessment due to student fatigue and lack of concentration during rating. They propose a practical limit of 20 concepts as a reasonable compromise between theoretical (validity, coverage of the domain) and practical (fatigue, time constraints) considerations.

The selected concepts are randomly combined and presented in $n(n-1)/2$ pairs (in which n is the number of concepts) to the learner who has to rate these on a 5, 7 or 9 point Likert scale. In the case where concepts are assessed for their relatedness, a 1 indicates that concepts are less related, while the 5, 7 or 9 indicate a high relation between concepts. For example, the relatedness rating of the concepts 'water' and 'ice' is likely to be higher (in the direction of 5, 7 or 9) than the relatedness rating of 'water' and 'iron' (more in the direction of 1). Ultimately, the pair ratings result in a matrix of proximity values between concepts indicating how closely the concepts are related (see Table 1).

Table 1 A set of hypothetical proximity data based on relatedness ratings between five concepts

	A	B	C	D	E
A	0	1	3	2	3
B	1	0	1	4	6
C	3	1	0	5	5
D	2	4	5	0	4
E	3	6	5	4	0

Note: The data in the matrix may be in the form of similarities, dissimilarities, probabilities, distances, coordinates, or features. The proximity values in the table are not the relatedness ratings on the Likert scale, but their transformation into distances. Smaller numbers represent pairs that are closely related. The distance between A–B is 1 which means that they are very closely related

(2) Knowledge representation

The next step is the *representation of the elicited knowledge*. One could say that the proximity matrix provides a representation of the knowledge of a domain or task, but the raw proximities in the matrix may be difficult to interpret. Therefore a scaling procedure has to be conducted to obtain a better representation of the underlying knowledge organization. Multidimensional scaling (MDS) can be used to represent semantic dimensions and the arrangement of concepts in a dimensional space (Gonzalvo et al. 1994). The Pathfinder procedure, on the other hand, uses a graph-theoretic distance technique (it uses the distance measured by the minimum number of links connecting two nodes in a graph) to represent the proximity matrix in a graphical network structure which can be either directed or undirected (Schvaneveldt et al. 1985). In Pathfinder the structural representation of the proximity values is referred to as a PFnet. Pathfinder represents concepts as nodes and distances as links in a network structure.

As a starting point Pathfinder links all concepts and assigns a weigh to each link (based on the ratings). Next, the Pathfinder algorithm removes direct links if there exists a shorter, indirect path that connects both concepts. This shorter, indirect path can be defined as a path with a lower link weight sums for the constituent links. For example, if the link weight of concept pair A–C is 6, but the sum of link weights between A–B and A–C is 4, then the direct link A–C is removed (Acton et al. 1994). Figure 1 shows the network structure after the application of this procedure on the proximity matrix of Table 1.

First, concept pairs with the shortest distance are processed. The shortest possible distance is 1. Table 1 shows there are two pairs (A–B and B–C) with a distance of 1. These concept pairs are directly connected because no shorter, indirect path exists (there is no lower link weight sum than 1). Then, distances of 2 are taken into account (A–D). A direct

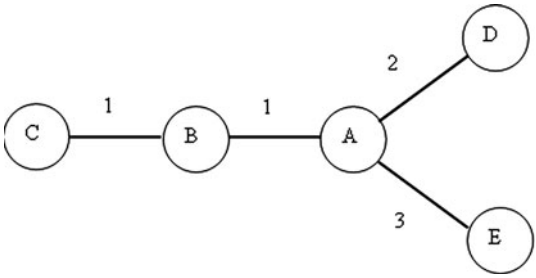


Fig. 1 Pathfinder network (PFnet) based on hypothetical proximity data in Table 1

link connects the concept pair A–D because no shorter path can be found. As an example, consider the potential shorter path in which A–D is connected via A–B and B–D. A–B has a link weighed sum of 1, B–D one of 4 (see Table 1) which makes 5 which is not a shorter path. The pairs A–E and A–C have distances of 3. The procedure creates a direct link for the pair A–E, but not for A–C. In the latter case a shorter indirect path from A to C exists via B (A–C and B–C both with a distance of 1 and thus with a shorter link weight of 2).

(3) Knowledge evaluation

The third step concerns the *evaluation of the knowledge representation* with a referent knowledge structure. Referent knowledge structures can be derived from an instructor, an expert other than the instructor or an analysis of instructional material. Before choosing a referent knowledge structure several issues have to be considered. The foremost consideration is whether a referent knowledge structure can be used in a domain or that a referent free structural assessment is more apt. Keppens and Hay (2008) have argued that 'hard' disciplines such as physics enable the classifications of relations between concepts into correct and incorrect ones. Consequently, they suggest that for these domains a referent based structural assessment with similarity scores can be used. For 'soft' domains (e.g., history) in which there is little agreement on the core concept and their relations they suggest a referent free structural assessment based on the internal consistency of the knowledge structure. For this purpose Pathfinder provides the coherence measure which assumes that if two concepts have similar relationships with other concepts, then the two concepts should be similar to one another. The coherence measure computes an indirect measure of similarity by correlating the ratings given for each item in a pair with all of the other concepts (e.g., with four concepts ABCD the pair AB would be correlated with the ratings of AC, AD, BC and BD) which gives the indirect similarity of AB—the extent to which A and B have similar relationships with other concepts. If we do this for all pairs, we can construct a half-matrix of indirect similarities. The coherence is the correlation between these indirect measures and the original ratings given for each pair of items. A more consistent set of ratings is supposed to yield a higher coherence (see also http://interlinkinc.net/FAQ.html#Coherence). It is suggested that high coherence is associated with a high level of expertise (cf. Dorsey et al. 1999). A second consideration involves the use of one or more experts to generate a referent knowledge structure. It has been suggested that each expert organizes domain knowledge differently (cf. Sternberg 1989) and that using multiple experts may introduce variability in the judgments of concept-pair relations which in turn may compromise the validity of the referent knowledge structure based on these experts. In this respect a study of Acton et al. (1994) comparing several types of referent structures may be clarifying. Among others they used an averaged expert referent structure consisting of six experts and it appeared that similarity between these experts was only .31. Although this 'inter-expert' similarity may seem low in an absolute sense, this averaged expert-based referent structure appeared to be superior for evaluating the knowledge structures of learners compared to referent structures based on a single expert or a consensus between experts (Acton et al. 1994; see also Day et al. 2001). This suggests that experts do not have to agree highly in an absolute sense for averaging expert ratings. It may also suggest that the variability causing the low 'inter-expert' similarity reflects random error rather than systematic differences in thinking which can be eliminated by averaging the structures (Acton et al. 1994; Day et al. 2001; Trumpower et al. 2010). Thirdly, the question can be raised whether one referent knowledge structure is sufficient. Some scholars have opposed that expertise may develop in stages in which

each stage has a specific qualitative knowledge organization (Acton et al. 1994; Boshuizen and Schmidt 1992; Kraiger et al. 1995; Shuell 1990). For example, in clinical reasoning Boshuizen and Schmidt adopt a 3-stage model consisting of respectively acquiring knowledge, practical experience and the integration of both types of knowledge. The knowledge representations during these stages are qualitatively different, for example, in the first stage novices depend on knowledge that they have to activate intentionally, while in the third stage experts have encapsulated knowledge and experience into clusters of diseases and symptoms which are activated automatically. This view suggests that there is not one 'ideal' referent structure but that each stage of development may require a different referent structure.

In structural assessment with Pathfinder the comparison of knowledge structures of learners with a referent structure can be conducted in several ways. In this paper we will focus on the *similarity between knowledge structures* and *graph-theoretic indices of the focal node*. The similarity between two knowledge structures (networks) can be defined as the number of common links between both networks divided by the total number of links. For example, consider two networks, each with four nodes (A, B, C, and D). Network 1 has three links (A–B, A–C, and A–D), and network 2 has three links (A–B, A–D, and B–C). The similarity for these two networks would be .50, the number of links the networks have in common (two: A–B and A–D) divided by four (the four links in both networks combined). The value of the similarity measure can range from 0 to 1, with 1 representing perfect similarity. Goldsmith and his colleagues (1990, 1991) originally developed a measure of network similarity which they called C (for Closeness). Unfortunately, the probability distribution for this measure has not been discovered. The similarity measure used in Pathfinder follows the hypergeometric probability distribution and thus computes values from this distribution to provide information concerning the similarity expected by chance. In our analysis we will use the Pathfinder similarity measure.

Because the similarity between the knowledge structures of learners and experts is assumed to reflect the degree of understanding a domain, a high level of similarity indicates better understanding of a domain or task than low similarity. Indeed a high level of similarity has proven to be more predictive for post-training performance than a low level of similarity in a variety of domains such as performance on a naval decision task (Kraiger et al. 1995), transfer after learning a complex skill in a video game (Day et al. 2001) and exam performance in statistics (Goldsmith et al. 1991).

Beside the similarity score researchers also have suggested to analyze and compare the central concepts in the knowledge structures and whether instruction will yield changes in what is regarded as the central concepts (Dayton et al. 1990; Nash et al. 2006). In order to analyze the central concepts these researchers propose three indices of a focal node: the highest degree node, the median node and the center node.

The highest degree node is the node with the greatest number of links (e.g., in Fig. 1 'A' is the highest degree node because it has three links whereas the other concepts only have one link). The median and center nodes are the least distant from all other nodes, but in a different way. To compute the median and center node a path distance matrix has to be created which contains the shortest path (i.e., the minimum number of links) for each pair of nodes. For the PFnet in Fig. 1, for example, a 5 by 5 nodes path distance matrix can be created. The minimum number of links between node 'C' and 'A' is 2 which is then inserted at the intersection of 'A' and 'C' in the matrix. The median node concerns the smallest average distance to all other nodes. It can be computed by averaging the number of links within each column of the matrix and selecting the lowest average. The center node is the node with the smallest maximum distance to any other node and can be

computed by finding the largest number of links in each column and then select the column with the smallest number. The nodes (concepts) indicated by these three indices can be considered the most central in a network (knowledge structure).

The comparison of these central concepts (focal nodes) in different knowledge structures may reveal which concepts the knowledge structures have in common. This method can be valuable because it makes a shift in understanding more concrete. For example, Nash et al. (2006) found that teachers regarded the concept 'inheritance', which is a fundamental concept in object oriented programming, to be more prominent after a training (it was the highest degree node and the median node), while before the training the concept was no focal node.

We have described the consecutive steps that have to be undertaken to conduct a structural assessment and the considerations and the pitfalls that have to be taken into account. We will now apply structural assessment in a learning environment with the game Code Red: Triage.

Learning a triage with a serious game

Triage

A primary triage (also called sieve) is the process of prioritizing patients based on the severity of their condition so as to treat as many as possible when resources are insufficient for all to be treated immediately. In the Netherlands, well defined procedures and protocols exist for the primary triage (cf. Hustinx et al. 2002). Figure 2 provides an overview of this procedure.

Four categories exist to which a victim can be assigned. T1 implies that immediate help is required; T2 implies that the injuries require urgent help; T3 implies that the victim is wounded but that further treatment can wait; although only a qualified physician is allowed to declare someone dead, a well-trained emergency responder may ascertain death and allocate someone to the category Dead. Although the procedure seems straightforward, the correct application depends on the condition of the victim (e.g., when a victim has an obstructed airway the procedure will be different from the situation in which a victim can breathe freely), the age of the victim (for children adapted values for breathing frequency and blood circulation are required) and environmental parameters (i.e., the outside temperature and the visibility).

Two arguments underpin our conclusion that structural assessment is appropriate for procedural domains such as a triage. First of all, triage is more a 'hard' discipline than a 'soft' discipline: there is substantial agreement about the important concepts and relations between concepts can be classified into correct and incorrect ones. Therefore a referent-based structural assessment seems appropriate. Second, we have stated that the prevailing view on knowledge structures is that they are associative networks of ideas, concepts, procedures and other forms of knowledge. The triage involves procedural knowledge (e.g., knowing when you have to check the breathing frequency), conceptual knowledge (e.g., how you can measure the breathing frequency), and situational knowledge (e.g., knowing that incidents with explosions cause specific types of injuries). This implies that the different entities in the triage connected with these types of knowledge are likely to be part of the knowledge structure. Moreover, these entities exhibit a pattern of relationship, for example, the entity 'Check breathing frequency' has a relation with the entity 'Pulse' because the pulse can be used to get an indication of the breathing frequency.

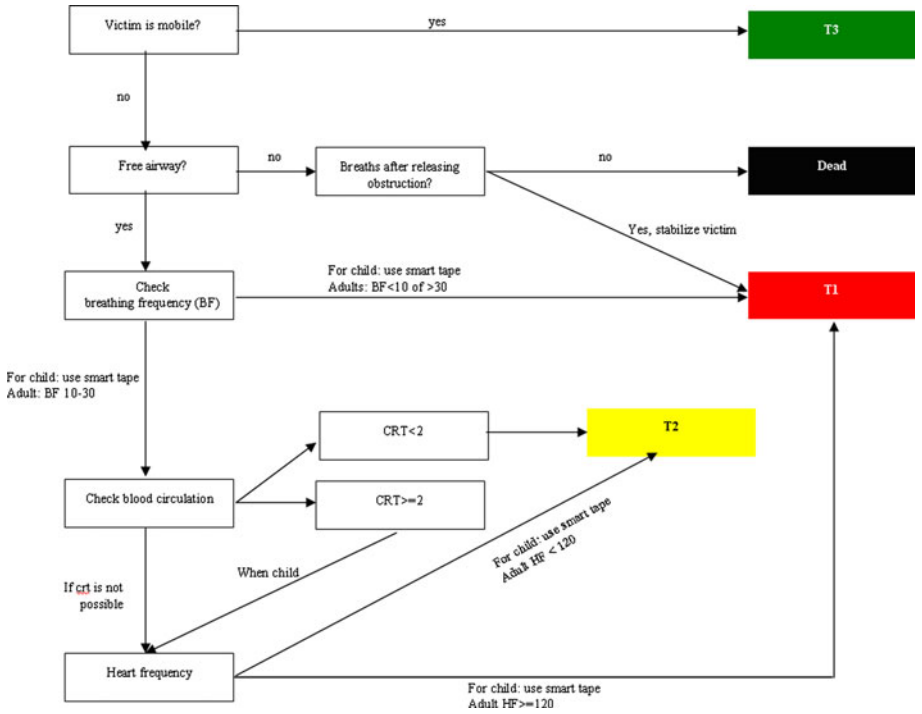


Fig. 2 The triage procedure

Code Red: Triage

It is obvious that it is crucial that medical responders learn the procedure and how to apply this in situations that resemble the conditions they have to face in reality, with factors such as time pressure, multimodal information sources (e.g., visual, auditory, tactile), chaos and emotions. Learning by book will not provide these situations and real life simulations are expensive and difficult to organize. The current state of software technology makes it possible to create games that meet these conditions.

The game *Code Red: Triage* is based on the Half Life 2 engine. The back story of the game is a bomb explosion in a subway station with 19 victims, 17 on the platform and two in the train. The participant, in the role of medical officer, arrives first at the location and has to perform a primary triage. During the gameplay the player learns the procedure and how to apply it to victims in different situations. The time for this task is 15 min which starts when the player encounters the first victim. In cooperation with triage experts a scenario was developed with 19 victims addressing the circumstances and injuries associated with subway bomb explosions.

The player starts the game in the station hall. Here the player is briefed about the situation and the task that has to be accomplished. From the station hall the player has to navigate to the platform where the victims lie. When a player approaches a victim an interface with nine buttons can be evoked by pressing the ‘e’ button on the keyboard (see Fig. 3).

With each button, information regarding the subject of the button can be evoked. For example, pressing the button ‘Ademweg’ (Dutch for Airway) displays information regarding the condition of the airway (‘is it blocked?’, ‘is the victim coughing?’ etc.). The

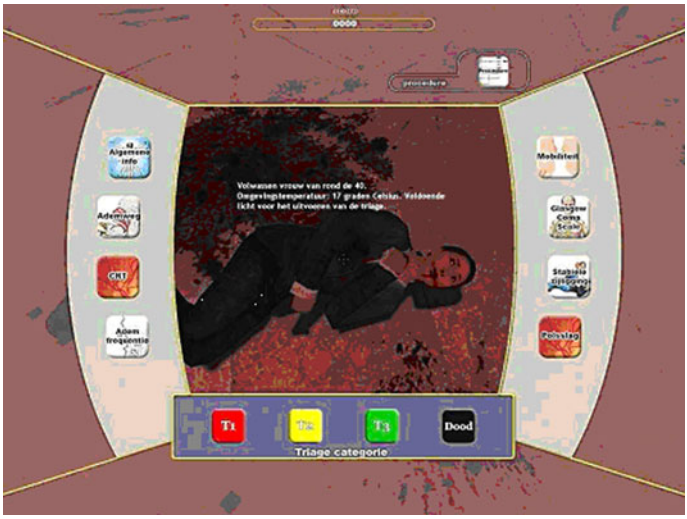


Fig. 3 Interface for triage in Code Red: Triage

player has to determine which of the buttons to use (the information required to adequately classify the victim) and the order in which these buttons have to be used (the appropriate procedure steps). The player can classify a victim by selecting one of the four buttons at the bottom of the screen that correspond to the four categories that are used for triage in the Netherlands. Game performance (fast and correct classification of victims) is reflected in a score on the screen. A correct classification yields 100 points. A penalty is subtracted from the score when the player takes longer than a preset time for each victim. See Van der Spek et al. (in press) for a more extensive description of Code Red: Triage.

Case study: Application of structural assessment on learning with Code Red: Triage

Participants

In the case study 19 participants volunteered. Nine of them were novices (mean age was 31.11, $SD = 14.21$) and 10 were advanced learners (mean age of 41.50, $SD = 9.13$). Novices were students and participants who responded on calls on internet and notice boards. They had little medical knowledge and no knowledge of the triage procedure they had to learn. The advanced learners were ambulance personnel recruited from two regional organizations involved in medical assistance. They had medical knowledge, but were not familiar with the triage procedure they had to learn.

Measurements

Traditional verbal assessment

A knowledge test consisting of 10 multiple choice items on declarative and procedural knowledge was used (Cronbach's alpha = .70). An example of a declarative knowledge item is:

Each triage category has its own color. The color for triage category 2 is:

a. Blue
b. Red
c. Green
d. Yellow

An example of a procedural knowledge item is:
During the primary triage of a victim you first have to examine:

a. Whether the airway is blocked
b. Whether the victim can walk
c. Whether there are external bleedings
d. The heart rate

Assessment of the knowledge structure

The concepts were obtained from a domain analysis by the authors using instructional material (cf. Hustinx et al. 2002) and the booklet 'De Nederlandse slachtofferregistratiekaart (The Dutch victim registration card)'. The list of concepts was adjusted after a review of and comments from two triage experts (an instructor and a trauma physician at a hospital). The final list of 13 concepts is presented in Table 2. All concepts were addressed in the game.

We used the rating program of the Pathfinder software with a 9-point Likert scale (1 set to highly unrelated and 9 set to highly related). Before the rating started the participants received some explanations and examples of how to rate concept pairs. These examples were not related to primary triage but involved general knowledge (e.g., high relationship: 'ice-water' versus low relationship: 'water-iron'). In addition, the participants were explicitly asked not to think too long about the relatedness of the concepts, but to rely on their first impression.

Three instructors involved in the triage training of medical personnel were used to calculate the expert-based referent knowledge structure. Each expert rated a total of 78

Table 2 Domain concepts used for structural assessment with associative networks

Domain concepts
Mobility
Capillary Refill Time (CRT)—nail bed
Capillary Refill Time (CRT)—forehead
Pulse
Cold weather
Triage Tape
Classify victim
Length victim
Respiratory rate
Airway
Stabilize victim
Blood circulation
Darkness

pairs ($n(n-1)/2$ with $n = 13$). Since the experts were used as referent and did not engage in a training session with Code Red: Triage, they only rated the pairs once.

The student knowledge structure was also determined by having each participant rate the 78 pairs. Because we wanted to compare the knowledge structures before and after the game, participants were asked to rate the 78 pairs twice. In general, the participants used approximately 10 min for rating the pairs. The reactions of participants also revealed that especially the second rating session, after playing the game, was strenuous because of fatigue and a decrease in concentration.

With Pathfinder software the relatedness ratings of the 78 pairs were transformed into structural graphical representations (PFnets). For experts only one PFnet was generated which was based on the averaged ratings of the three experts in the Knowledge elicitation stage. For each participant two PFnets were created: one of the pair ratings before the game and one of the pair ratings after the game. As an example Fig. 4 shows a PFnet of the averaged ratings of the experts.

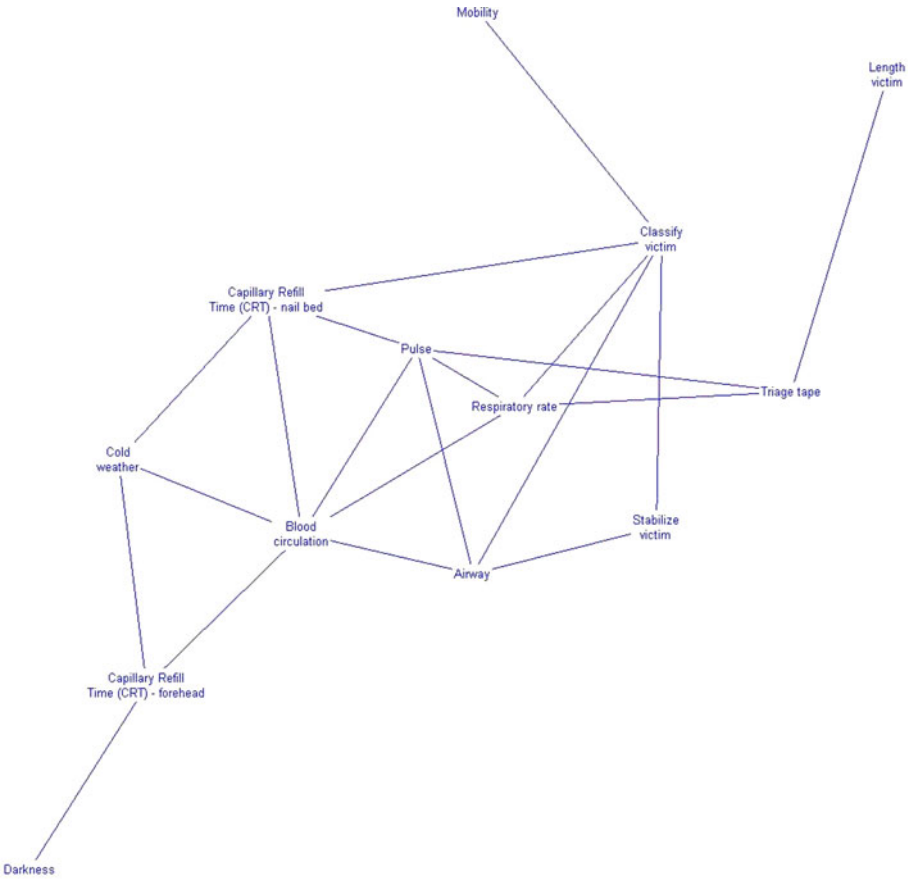


Fig. 4 PFnet of the experts. The expert PFnet has 21 links. A visual inspection tells that three concepts 'Classify victim', 'Pulse', and 'Blood circulation' have the most connections (5 or more) with other concepts. Three concepts are relatively isolated (having only one connection with another concept)

Table 3 Means and standard deviations (between brackets) for the knowledge test (all items, declarative items, and procedural items), similarity scores and coherence scores before and after game

	Novices ($N = 9$)		Advanced learners ($N = 10$)	
	Before game	After game	Before game	After game
Knowledge test				
All items	4.00 (1.22)	6.89 (1.76)	7.70 (1.16)	8.30 (1.16)
Declarative items	2.89 (.93)	3.56 (.53)	3.90 (.74)	4.30 (.67)
Procedural items	1.11 (.78)	3.33 (1.50)	3.80 (.92)	4.00 (1.15)
Structural assessment				
Similarity scores	.11 (.07)	.16 (.07)	.24 (.13)	.21 (.08)
Coherence scores	.50 (.12)	.31 (.24)	.49 (.24)	.30 (.20)

Results

Table 3 shows the mean scores on the knowledge tests (for all items and split in five declarative and five procedural items), the similarity scores and the coherence scores. Although we assumed a referent-based evaluation of knowledge structures suits the triage domain we also included the coherence scores in order to have a full coverage of the Pathfinder procedure.

Traditional verbal assessment

When we consider all items the analysis of the knowledge tests show that both novices and advanced learners perform better on the posttest compared to the pretest (advanced: $Z = -2.121$, $p < .05$; novices: $Z = -2.230$, $p < .05$). A closer look reveals that on the procedural items novices perform better on the posttest compared to the pretest ($Z = -2.448$, $p < .05$), while this is not true for the declarative items ($Z = -1.561$, $p > .05$). For advanced learners the pattern is reversed: on declarative items they perform better on the posttest compared to the pretest ($Z = -2.000$, $p < .05$), while this is not true for the procedural items ($Z = -.816$, $p > .05$). Considering all items the analysis also makes clear that before the game advanced learners performed much better than novices ($Z = -3.560$, $p < .001$), but not after the game ($Z = -1.804$, $p > .05$). A closer look shows that before the game advanced learners indeed performed better than novices on both declarative and procedural items (resp. $Z = -2.196$, $p < .05$, $Z = -3.621$, $p < .001$), while after the game the advanced learners only perform better on declarative but not on procedural items (resp. $Z = -2.287$, $p < .05$, $Z = -.935$, $p > .05$).

Structural assessment: similarity and coherence scores

The similarity between the three experts was .41 which we assumed to be large enough to use an averaged expert referent structure (see also Acton et al. 1994). Initially, the similarity with the referent knowledge structure was larger for advanced learners compared to novices ($Z = -2.265$, $p < .05$). After the game the larger similarity for advanced learners had disappeared ($Z = -1.192$, $p > .05$). We also conducted related-samples Wilcoxon tests to test whether playing the game would yield an increase in the similarity with the referent knowledge structure. The analysis revealed a significant increase in similarity for

novices ($Z = -1.96$, $p < .05$), but not for advanced learners ($Z = -.78$, $p > .05$). The coherence scores for novices did not differ from the advanced learners neither before ($Z = 0$, $p > .05$) nor after the game ($Z = -.132$, $p > .05$). For novices the decrease in coherence after the game was weakly significant ($Z = -1.82$, p was .069), while the decrease was significant for the advanced learners ($Z = -2.19$, $p < .05$).

Structural assessment: PFnets

Figure 5a, b show the PFnets of the knowledge structure of the novices before and after they played the game (for this purpose the novices' knowledge structures were averaged). Likewise Fig. 6a, b depict the averaged PFnets for the advanced learners before and after they played the game.

Next the PFnets (including the expert's, see Fig. 4) were used to compute the three focal nodes: highest degree node, median node and center node. For both the median and center node a path distance matrix was created (in this case 13 by 13 nodes). Figure 7 shows the path distance matrix for the averaged expert PFnet.

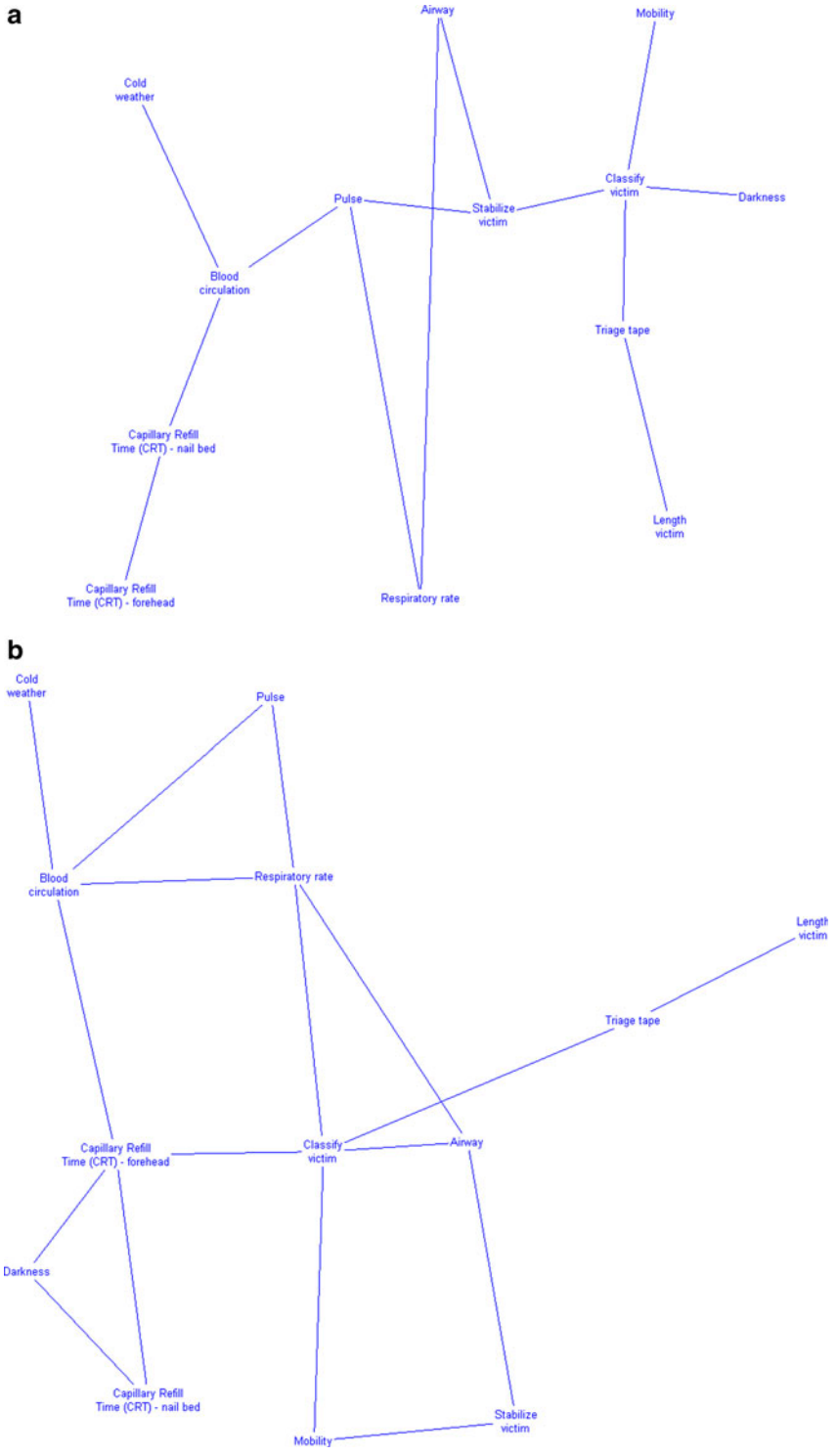
Table 4 shows the focal indices for learning triage with Code Red: Triage.

We first compared the most central concepts in the PFnets. We assume that the expert's central concepts ('Circulation', 'Pulse', 'Breathing') can be regarded as critical information that guide a quick and effective triage. Initially novices only share the central concept 'Pulse' with the experts. After the training it seems that novices emphasize 'Classify victim' (both highest degree node and median node). Advanced learners show another picture. Before the game only one central concept was shared with the experts ('Circulation'), while after the game the central concepts of experts and advanced learners converged (although distributed differently across the three indices).

Conclusion and discussion

In educational research verbal items are often used to assess the effectiveness of learning environments such as serious games. An alternative is the use of structural assessment in which the organization of knowledge is assessed by having individuals rate the relatedness of pairs of concepts. In this paper we showcase the Pathfinder model of structural assessment consisting of knowledge elicitation, knowledge representation and knowledge evaluation on the serious game Code Red: Triage. In addition, we compare novices and advanced learners performance on traditional and structural assessment.

For novices the traditional verbal assessment shows that they learn form playing the game, but this was largely attributable to better procedural knowledge and not to an increase in declarative knowledge. Regarding structural assessment the similarity measure shows that their knowledge structures become more similar with the expert referent structure which is in line with the improvement measured by traditional verbal assessment when all items are taken into account. However, the similarity measure is not able to discern between improvement on procedural and declarative knowledge. The coherence measure shows a decrease after training with the game. The analysis of focal nodes shows that the concepts that novices regard as central in the triage are divergent from what experts regard as central concepts before as well as after playing the game. It is remarkable that the PFnets have many links both before and after the training, while Pathfinder was parameterized to construct a PFnet with $n - 1$ links. For advanced learners the traditional verbal assessment shows that they learn from playing the game, but in contrast to novices

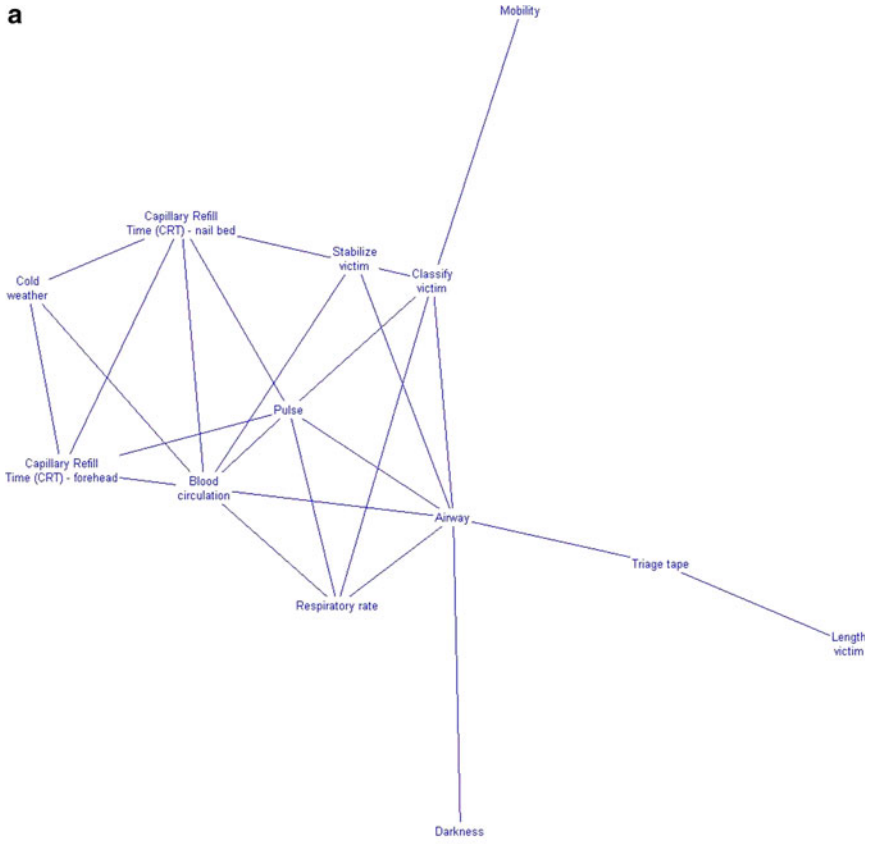

a
Airway
Mobility
Cold weather
Classify victim
Darkness
Pulse
Stabilize victim
Blood circulation
Triage tape
Capillary Refill Time (CRT) - nail bed
Length victim
Capillary Refill Time (CRT) - forehead
Respiratory rate
b
Cold weather
Pulse
Blood circulation
Respiratory rate
Length victim
Triage tape
Capillary Refill Time (CRT) - forehead
Classify victim
Airway
Darkness
Capillary Refill Time (CRT) - nail bed
Stabilize victim
Mobility

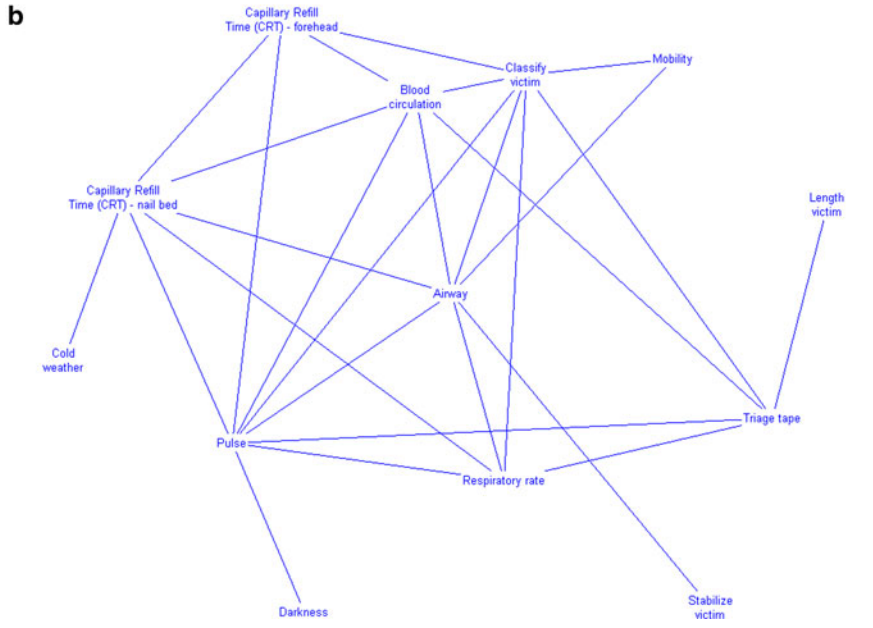
◀ **Fig. 5** **a** Averaged PFnets of novices before the game. The novices PFnet before the game has relatively few links (13). Also the visual inspection indicates there is no concept with many links to other concepts ('Classify victim' may be an exception). Also notable s that five concepts are relatively isolated (having only one connection with another concept). **b** Averaged PFnets of novices after the game. After the game the novices PFnet has become more branched (less isolated concepts with only one connection with another concept) with 25 links and more concepts with many links to other concepts. Two concepts are relatively isolated (having only one connection with another concept)

this was largely attributable to better declarative knowledge and not to an increase in procedural knowledge. Regarding structural assessment the similarity measurement shows that the knowledge structures did not become more similar with the expert referent structure after training with a computer game. The coherence measure shows a decrease after training with the game. The analysis of focal nodes on the other hand showed that advanced learners and experts regarded the same concepts as central after playing the game, whereas this was not the case before the game. Also the PFnets of the advanced learners have many links both before and after the training, while Pathfinder was parameterized to construct a PFnet with $n - 1$ links.

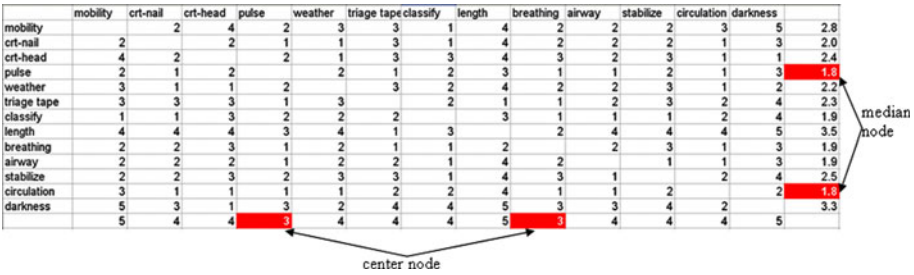
The pattern of results yields at least three questions that should be discussed. First of all, we failed to find an increase in similarity between the advanced learners and the expert referent after the game. The aforementioned stage view on expertise development (Acton et al. 1994; Boshuizen and Schmidt 1992; Kraiger et al. 1995; Shuell 1990) may provide an explanation. The stage view contends that there is no 'ideal' referent structure, but that each level of expertise requires a different referent structure. Our expert referent structure was based on concepts originating from an analysis of instructional material. These concepts reflect a theoretical account of the triage procedure which may not align well with the concepts and knowledge organization of advanced learners who already have some practical experience. It is possible that a different referent structure which is based on concepts directly originating from the experts who also have practical experience would be more in line with the way advanced learners organize their knowledge. If this stage view plays a role, it is plausible that novices may show a larger increase in similarity with advanced learners as a referent than with experts as a referent. Indeed novices showed a larger increase in similarity with advanced learners as referent (the increase in similarity was $.18 - .09 = .09$) than with an expert referent (increase was $.16 - .11 = .05$). The second question concerns the high number of links in the PFnets. Although such a high number is not unique (cf. Gonzalvo et al. 1994), it is surprising because Pathfinder was set to construct a minimally connected average PFnet. A possible explanation is that participants have used many extreme relatedness ratings (so many 1s and 9s) or rated many pairs as highly related (8s and 9s). A further analysis of the proximity data of novices showed that after the training some participants tend to give more extreme relatedness ratings than before the training (see Fig. 8). Note that most novices show an increase in extreme relatedness ratings after the training. Extreme relatedness ratings results in a large number of tied values, because participants only utilize a small portion of the scale. The large number of links can be ascribed to these extreme relatedness ratings because Pathfinder includes all links when there are tied distances.

The third question pertains to the decrease in coherence. If it is true that the coherence of knowledge structures will become larger with increasing expertise, the observed decrease of the coherence for both novices and advanced learners after training with the game is remarkable. It is possible that the lower coherence can be partly attributed to the fatigue and decreasing concentration during the second rating. The participants had learned


a
Mobility
Capillary Refill
Time (CRT) - nail bed
Stabilize
victim
Classify
victim
Cold
weather
Pulse
Capillary Refill
Time (CRT) - forehead
Blood
circulation
Airway
Triage tape
Respiratory rate
Length
victim
Darkness


b
Capillary Refill
Time (CRT) - forehead
Classify
victim
Mobility
Blood
circulation
Capillary Refill
Time (CRT) - nail bed
Length
victim
Airway
Cold
weather
Triage tape
Pulse
Respiratory rate
Darkness
Stabilize
victim

◀ **Fig. 6** **a** Averaged PFnets of advanced learners before the game. Before the game the advanced learners PFnet is dense with 24 links. Many concepts have five or more connections with other concepts. Three concepts are relatively isolated (having only one connection with another concept). **b** Averaged PFnets of advanced learners after the game. After the game the advanced learners PFnet remains dense with 25 links. The visual inspection shows an increase in the number of connections of concepts with other concepts (e.g., 'Classify victim' increased from five links to seven). Four concepts are relatively isolated (having only one connection with another concept)



	mobility	crt-nail	crt-head	pulse	weather	triage tape	classify	length	breathing	airway	stabilize	circulation	darkness	
mobility		2	4	2	3	3	1	4	2	2	2	3	5	2.8
crt-nail	2		2	1	1	3	1	4	2	2	2	1	3	2.0
crt-head	4	2		2	1	3	3	4	3	2	3	1	1	2.4
pulse	2	1	2		2	1	2	3	1	1	2	1	3	1.8
weather	3	1	1	2		3	2	4	2	2	3	1	2	2.2
triage tape	3	3	3	1	3		2	1	1	2	3	2	4	2.3
classify	1	1	3	2	2	2		3	1	1	1	2	4	1.9
length	4	4	4	3	4	1	3		2	4	4	4	5	3.5
breathing	2	2	3	1	2	1	1	2		2	3	1	3	1.9
airway	2	2	2	1	2	2	1	4	2		1	1	3	1.9
stabilize	2	2	3	2	3	3	1	4	3	1		2	4	2.5
circulation	3	1	1	1	1	2	2	4	1	1	2		2	1.8
darkness	5	3	1	3	2	4	4	5	3	3	4	2		3.3
	5	4	4	3	4	4	4	5	3	4	4	4	5	

Fig. 7 Path distance matrix for the averaged expert PFnet

Table 4 Three graph theory indices of focal node for PFnets in Code Red: Triage

	Highest degree node[a]	Median node[b]	Center node[c]
Novices			
Before the game	Classify 4	Stabilize 2.2	Pulse, stabilize 4
After the game	Classify 5	Classify 1.7	Breathing, airway 3
Advanced learners			
Before the game	Circulation, airway 7	Airway 1.4	Airway 2
After the game	Pulse 8	Pulse 1.3	Pulse, breathing, circulation 2
Experts	Circulation 6	Pulse, circulation 1.8	Pulse, breathing 3

[a] The node(s) with the greatest number of links. The highest degree is given below the node names

[b] The node(s) with the smallest average distance to all other nodes. The smallest average distance is given below the node names

[c] The node(s) with the smallest maximum distance to any other node. The smallest maximum distance is given below the node names

from the game, but they were less precise in eliciting the underlying knowledge structure. An alternative explanation may be that the knowledge organization became more differentiated in connected clusters. The relation within these clusters may have been more cohesive, but they suppress the cohesiveness of the knowledge structure as a whole. For example, a cluster around physical body characteristics versus a cluster around environmental, external characteristics.

Altogether the results suggest that structural assessment measures an individual's understanding of a domain at least differently from verbal assessment. The advantage of verbal assessment is that it provides a more nuanced picture regarding declarative and procedural knowledge. Structural assessment adds an in-depth understanding of the concepts that are regarded important by learners (and how these deviate from the concepts that

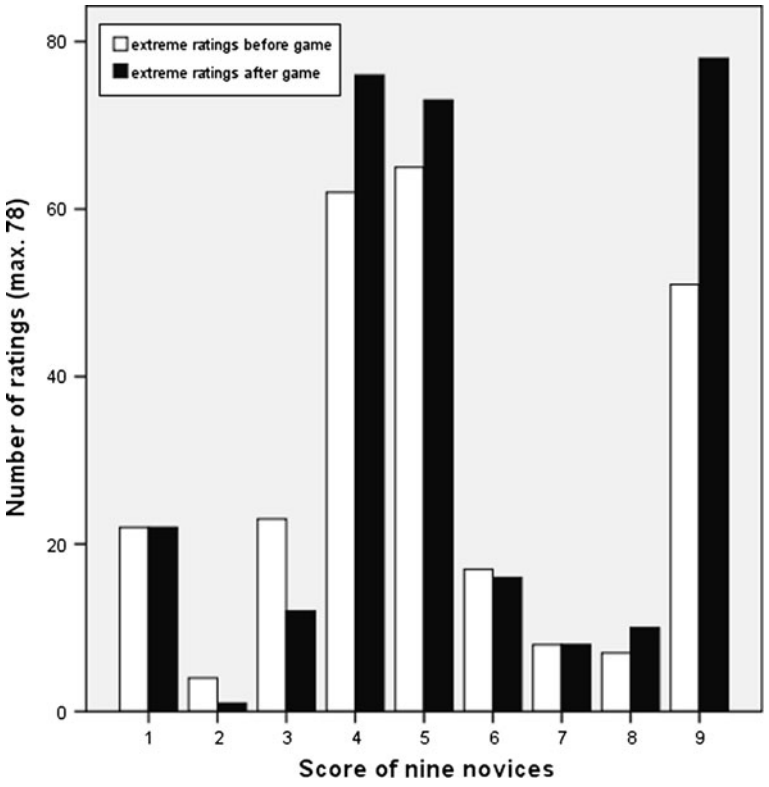


Fig. 8 Number of extreme ratings (1–9) for novices before and after the game

are actually important). The results of the similarity measure suggest that one referent structure may not be sufficient for several levels of expertise. Another argument that may further substantiate the potential of structural assessment in serious games follows from research regarding the predictive validity of similarity scores. Although we were not able to compare the similarity scores with a valid post-training measure (e.g., performance during physical emergency simulations with actors playing victims), other studies have shown that similarity scores can be predictive for post-training performance (Day et al. 2001; Goldsmith et al. 1991; Kraiger et al. 1995).

It is obvious that the results should be viewed in the light of some clear limitations that adhere to this study. To start with, the number of participants was low. We need to replicate studies like this with more participants. A second caveat is that the instruction time in the case study was quite short (15 min). It can be argued that such a short instruction time is not long enough to effectuate sustainable cognitive effects. Indeed, an analysis of the instruction or practice time in some structural assessment studies shows much longer periods varying from a semester in statistics (Acton et al. 1994), a 16 week course on psychological research techniques (Goldsmith et al. 1991), a three week course in physiology (McGaghie et al. 2000), an 80 h computer programming workshop (Nash et al. 2006) to a three day training in a computer game (Day et al. 2001). However, structural assessment has also been reported to be effective as well with training times less

than 60 min (cf. Rose et al. 2007 on accounting; Kraiger et al. 1995 on naval decision-making simulation).

Our exploration has also exposed some avenues for future research. To start with, more research is needed to use structural assessment beyond its common employment as a method for summative assessment (i.e., it reflects the degree in which a learner understands a domain/task). Trumpower et al. (2010) have argued that structural assessment can also be employed for formative assessment in the sense that it may also reveal the strengths and weaknesses of learners in a specific area of a domain or task. In their study participants learned about a computer programming language followed by structural assessment and several problem solving tasks. The researchers were able to connect two distinct types of problems with two distinct subsets of links. They found that the presence of these subsets in the PFnets of the participants differentially predicted performance on two types of problems. For example, participants with PFnets containing the subset of links about control structures in programming performed better on a problem solving task that required understanding of control structures than participants who lacked these links. The presence of this subset did not influence performance on problem solving tasks that were based on the other problem type. The ability of structural assessment to uncover specific strengths (e.g., control structure links in PFnet) and weaknesses (e.g., not having control structure links in PFnet) is a first step towards its formative use in learning and instruction. Secondly, more research is required regarding the appropriate referent structures for distinctive levels of expertise. Although research has been conducted on the effect of different referent structures (see Acton et al. 1994; Day et al. 2001), these studies have not made an explicit relation with different levels of expertise. Thirdly, scholars have argued that PFnets of referents can be used as an organizational structure, for example in the domain of hypertext (Jonassen 1988; McDonald et al. 1990) or as complement for a linearly organized syllabus for a course (Acton et al. 1994; Schvaneveldt et al. 1985). It would be interesting to investigate whether a PFnet of a referent structure can be effectively used as an instructional device. Fourth, we believe that the analysis of PFnets with focal nodes is a neglected dimension of structural assessment. We advocate research in this area. We also suggest more research in other techniques to analyze the focal nodes. For example, the highest degree node is the concept with the greatest number of links. The other concepts have a lower number of links and are therefore neglected by the researcher. What we propose is to consider all degree nodes of a knowledge structure and calculate correlations between the knowledge structures. If two knowledge structures agree on which nodes are high and which are low, the r value (correlation) will be larger, while less agreement will lower the r value.[1] In Appendix, we have applied this method to our data set.

To conclude, we will provide four guidelines based on the literature review and the case study results that may help to avoid some pitfalls: (1) Determine whether the domain allows a referent based structural assessment (agreement on important concepts in a domain) or that a referent free assessment is more appropriate. In the former situation we advise the Pathfinder similarity score, in the latter situation the coherence measure (also available in Pathfinder) can be used. For some domains such as mathematics or physics it may be clear that there is agreement on important concepts, but when this is uncertain in a domain we propose to determine a set of concepts, then have three or more experts (or

[1] We are grateful for this suggestion by one of the anonymous reviewers.

instructors) conduct the pair wise ratings and finally calculate the similarity among the experts. Very low similarity scores may indicate there is little agreement on the core concepts; (2) When a referent based structural assessment is chosen, consider carefully which type of referent is most suitable. This may be important when different levels of expertise are involved. Although many studies have found that averaged expert referent structures are superior in terms of predictive validity (Acton et al. 1994; Day et al. 2001), other types of referent structures may be more appropriate in specific situations (e.g. instructors instead of experts or the best students in a class). When different levels of expertise can be discerned in the data set also consider the use of higher levels of expertise as a referent for lower levels of expertise (as we did with advanced learners and novices). This may give some information whether the correct referent was used; (3) The concepts should be unambiguous and their number should not be too high. Although some participants found it difficult to maintain their concentration when they had to rate 13 concepts the second time, we still believe it is appropriate to stick to the maximum of 20 concepts (see Trumpower et al. 2010). Of course this also depends on issues like the domain, the specific task that is being trained and the mental impact of all measurements (if the intention is to measure only knowledge structures more concepts can be used compared to the situation in which also other measures such as pre- and posttests are used). In comprehensive domains in which it is likely one needs more than 20 concepts for an adequate representation, it can be an option to focus on a subset of the domain. For example, in one of our studies we found that structural assessment with a subset of the triage procedure (covered with eight concepts) was able to discern differences between participants who received cueing during Code Red: Triage and a group without cueing (van der Spek et al. 2010). Ambiguous concepts may jeopardize the quality of structural assessment. Consider a small pilot with experts and a sample from the target group to identify potential ambiguous concepts. For example, with the term 'stabilise victim' we meant the act of putting a person in a stable sideways position, something you do to relieve pressure on the airway. However during the expert interviews one expert thought 'stabilise victim' implied the stabilisation of the airway, breathing and circulation; a well-known medical treatment, but not part of the primary triage. The interpretation of this concept could be problematic for advanced learners because they already had general medical knowledge. This information helped us to adapt the instruction to the participants before they engaged in the structural assessment; (4) While the similarity score compares the knowledge structure as a whole with that of a referent, the more in-depth analysis of the PFnets (the graphical representations of the knowledge structures) can be used to obtain additional information about the knowledge structures. This analysis may involve visual inspection of PFnets or the calculation of focal node indices. In particular the latter technique seems valuable because it may reveal that persons with a certain level of expertise fail to recognize one or more important concepts in a domain (like the novices in this case study). In this way the calculation and analysis of focal nodes can be regarded as formative assessment, because it concretely reveals weaknesses in knowledge structures.

Acknowledgement This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie).

Appendix

We calculated the correlations between the PFnets in terms of degree nodes and created the table below with for each PFnet the number of links for each concept.

	Expert	Novices before game	Novices after game	Advanced learners before game	Advanced learners after game
Blood circulation	6	3	4	7	6
Classify victim	5	4	5	5	7
Pulse	5	3	2	6	8
Airway	4	2	3	7	7
CRT-nail bed	4	2	2	5	6
Respiratory rate	4	2	4	4	5
CRT-forehead	3	1	4	4	4
Triage tape	3	2	2	2	5
Cold weather	3	1	1	3	1
Stabilize victim	2	3	2	4	1
Mobility	1	1	2	1	2
Darkness	1	1	2	1	1
Length victim	1	1	1	1	1

The table shows for example that 'blood circulation' has six links with other nodes (and thus it is the highest degree node) in the average expert knowledge structure, while 'length' has a degree node of 1 meaning that it is only linked with one other concept (see also Fig. 4).

The table below shows the correlations between the PFnets.

	Expert	Novices before game	Novices after game	Advanced learners before game	Advanced learners after game
Expert					
Novices before game	.71**				
Novices after game	.61*	.53			
Advanced learners before game	.88**	.66*	.53		
Advanced learners after game	.87**	.64*	.57*	.80**	1

** Means significant at .01 level; * means significant at .05 level

The calculation of correlations between PFnets shows that experts and novices highly agree in the degree which concepts have a high number of links and which concepts have a low number of links. However, the correlations between experts and advanced learners seem higher suggesting that they very highly agree in the degree which concepts have a high number of links and which concepts have a low number of links with other concepts.

References

Acton, W. H., Johnson, P. J., & Goldsmith, T. E. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology, 86*, 303–311.

Adelson, B. (1981). Problem solving and the development of abstract categories in programming languages. *Memory and Cognition, 9*, 422–433.

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology, 57*, 204–213.

Boshuizen, H. P. A., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science, 16*, 153–184.

Day, E. A., Arthur, W., Jr., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology, 86*, 1022–1033.

Dayton, T., Durso, F. T., & Shepard, J. D. (1990). A measure of the knowledge reorganization underlying insight. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 267–277). Norwood, NJ: Ablex.

Diekhoff, G. M. (1983). Testing through relationship judgments. *Journal of Educational Psychology, 75*, 227–233.

Dorsey, D. W., Campbell, G. E., Foster, L. L., & Miles, D. E. (1999). Assessing knowledge structures: Relations with experience and posttraining performance. *Human Performance, 12*, 31–57.

Glaser, R., & Chi, M. T. (1989). Overview. In M. T. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv–xxviii). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldsmith, T. E., & Johnson, P. J. (1990). A structural assessment of classroom learning. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 241–254). Norwood, NJ: Ablex.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83*, 88–96.

Gonzalvo, P., Cañas, J. J., & Baja, M.-T. (1994). Structural representations in knowledge acquisition. *Journal of Educational Psychology, 86*, 601–618.

Hustinx, P., Meeuwis, D., & Hermans, R. (2002). *Geneeskundig management bij grootschalige incidenten (Major incident medical management and support: The practical approach)*. The Netherlands, Utrecht: De Tijdstroom.

Jonassen, D. H. (1988). Designing structured hypertext and structuring access to hypertext. *Educational Technology, 28*, 13–16.

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Keppens, J., & Hay, D. (2008). Concept map assessment for teaching computer programming. *Computer Science Education, 18*, 31–42.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*, 311–328.

Kraiger, K., Salas, E., & Cannon-Bowers, J. A. (1995). Measuring knowledge organization as a method for assessing learning during training. *Human Factors, 37*, 804–816.

McDonald, J. E., Paap, K. R., & McDonald, D. R. (1990). Hypertext perspectives: Using pathfinder to build hypertext systems. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 197–211). Norwood, NJ: Ablex.

McGaghie, W. C., McCrimmon, D. R., Mitchell, G., Thompson, J. A., & Ravitch, M. M. (2000). Quantative concept mapping in pulmonary physiology: Comparison of student and faculty knowledge structures. *Advances in Physiology Education, 23*, 72–81.

Messick, S. (1984). Abilities and knowledge in educational achievement testing: The assessment of dynamic cognitive structures. In B. S. Blake (Ed.), *Social and technical issues in testing: Implications for test construction and usage* (pp. 156–172). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nash, J. G., Bravaco, R. J., & Simonson, S. (2006). Assessing knowledge change in computer science. *Computer Science Education, 16*, 37–51.

Pirnay-Dummer, P., Ifenthaler, D., & Spector, J. (2010). Highly integrated model assessment technology and tools. *Educational Technology Research and Development, 58*, 3–18.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1985). *Pathfinder: Scaling with network structures*. Las Cruses: (Memorandum in Computer and Cognitive Science, MCCS-85-9, Computing Research Laboratory), New Mexico State University.

Shuell, T. J. (1990). Phases of meaningful learning. *Review of Educational Research, 60*, 531–547.

Sternberg, R. J. (1989). Domain-generality versus domain-specificity: The life and impending death of a false dichotomy. *Merill-Palmer Quarterly, 35*, 115–129.

Trumpower, D. L., Sharara, H., & Goldsmith, T. E. (2010). Specificity of structural assessment of knowledge. *The Journal of Technology, Learning, and Assessment, 8*(5), 1–31. Retrieved April 8, 2010 from http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1238&context=jtla.

van der Spek, E. D., van Oostendorp, H., Wouters, P., & Aarnoudse, L. (2010). Attentional cueing in serious games. In K. Debattista, M. Dickey, A. Proença, & L. P. Santos (Eds.), *2010 second international conference on games and virtual worlds for serious applications* (pp. 119–125). Los Alamitos, CA: IEEE.

van der Spek, E. D., Wouters, P., & van Oostendorp, H. (in press). Code Red: Triage or cognition-based design rules enhancing decionmaking training in a game environment. *British Journal of Educational Technology*.

Wouters, P., Van der Spek, E. D., & van Oostendorp, H. (2009). Current practices in serious game research: A review from a learning outcomes perspective. In T. M. Connolly, M. Stansfield, & L. Boyle (Eds.), *Games-based learning advancements for multisensory human computer interfaces: Techniques and effective practices* (pp. 232–255). Hershey, PA: IGI Global.

Pieter Wouters is researcher in the Cognition and Communication section within ICS. He holds a PhD (2007) in Instructional Design (How to optimize cognitive load for learning from animated models) from the Open University of the Netherlands. His current research focuses on cognitive and motivational processes in learning from serious games and game discourse analysis.

Erik D. van der Spek MSc is a PhD student in the Cognition and Communication section within the ICS. He graduated in Computer Science on the affective appraisal of virtual environments under the influence of cybersickness. In his current research, he aims to improve the effectiveness of serious games by cognitively engineering the game design and educational content, culminating in general design guidelines for serious games developers.

Herre van Oostendorp is Associate professor and head of the Cognition and Communication section within ICS. His background is (experimental) cognitive psychology. His teaching and research activities are on the domain of Human–Computer Interaction. He is a specialist on the areas of web navigation, hypertext comprehension, animation and usability engineering.