Some Observations Regarding the Assessment of Learning in Serious Games

Pieter Wouters, Erik van der Spek, Herre van Oostendorp
Utrecht University, Department of Information and Computing Sciences
Padualaan 14, Utrecht, The Netherlands
{pieterw, eriks, herre}@cs.uu.nl

Abstract. Verbal, visual and structural assessment methods were investigated with a serious game. Hypothesis 1 predicted that structural assessment (defined as the similarity between the knowledge structure of the player and those of experts) measures another pattern of learning results than verbal assessment. In study 1, game training yielded an increase in similarity (structural assessment) for novices, but not for advanced learners. This effect was not found with verbal tests. In study 2 both assessment methods revealed an increase after the game, but the effect size for structural assessment was larger. Hypothesis 2 predicting that gaming performance is better reflected in visual compared to verbal assessment was confirmed. The impact of the results on assessment of serious games is discussed.

Keywords: serious games; game-based learning; triage; structural assessment.

Introduction

Reviews suggest serious games are not always effective (cf. Wouters, van der Spek & van Oostendorp, 2009). Computer games are often contextual, complex and visually rich systems in which individuals may learn differently than learning written, verbal materials. This paper focuses on two differences. *Firstly*, players immersed in a game may have little opportunity for verbalization of, and reflection on their actions. This may yield implicit learning which is difficult to verbalize and thus difficult to assess with an assessment that is mainly verbal. Structural assessment assumes that someone's knowledge structure of a domain can be represented as a network of nodes and relations (Goldsmith, Johnson & Acton, 1991). This method is less verbal, more conceptually, and maybe more appropriate for game-based learning. Our first hypothesis contends that structural assessment measures another pattern of learning results than mainly verbal assessment. *Secondly*, computer games are often strongly visual and take place in a certain context. Our second hypothesis contends that performance during gaming is better reflected by performance on visual-oriented than mainly verbal assessment items.

Method

Participants and Design. For this explorative analysis we used data from two earlier pilot studies. Study 1 investigated the effect of expertise (N = 9 novices *vs.* N = 10 advanced learners). Study 2 investigated whether auditory cues could support players to focus on relevant information in the game (N = 11 auditory cues *vs.* N = 10 no cues). In both studies the game Code Red: Triage was used.

Code Red: Triage. In the game the players are confronted with an explosion in a subway with many casualties. The player has to navigate to the platform and conduct a triage (i.e., classify victims in one of four categories based on their injuries, see also Van der Spek *et al.*, in press).

Structural Assessment. A domain analysis yielded concepts that were presented in pairs (e.g., 'pulse' – 'respiratory rate') which participants had to rate on relatedness on a 9-point scale ranging from 'not at all related' to 'highly related'. In study 1, 13 concepts were used (78 pairs), study 2 used 8 concepts (26 pairs). Pathfinder software was used to calculate the similarity between the knowledge structures of the participants (based on the ratings of the pairs) and a referent expert knowledge structure (based

on ratings of three instructors). The assumption is that a high similarity with the expert structure reflects a better understanding of the domain (cf. Goldsmith *et al.*, 1991).

Verbal Assessment. Participants received 10 verbal, fully textual (multiple choice) questions measuring factual and procedural knowledge as a pretest. In the posttest version the questions were presented in a different order. After the game also 4 additional verbal questions (comparable with the visual assessment questions) were presented (see Figure 1, left).

Visual Assessment. After the game the participants received 4 mainly visual (multiple choice) questions (but still some verbal information was present). The visual questions resembled the screen presentation in the game (see Figure 1, right). All questions tapped the knowledge regarding the application of the triage procedure.

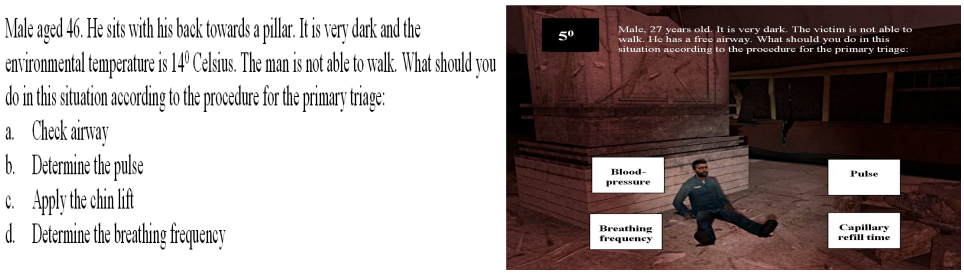

Figure 1. Example of a verbal (left) and visual (right) item.

Game performance. Game performance itself (fast and correct classification of victims) is reflected in a score on screen. A correct classification yielded 100 points. A penalty was subtracted from the score when the player took longer than a preset time for each victim.

Procedure. In both studies: (1) measuring structural assessment, (2) pretest (10 multiple choice, verbal items), (3) short explanation on triage, (4) the game (15 minutes), (5) measuring structural assessment, (6) posttest (pretest items in a different order) and (7) visual and additional verbal questions.

Results

Table 1: Means and Standard Deviations (between brackets) of Study 1 and 2.

	Verbal Assessment after - before	Structural Assessment after - before	Visual Assessment **(four items)**	Additional Verbal Assessment **(four items)**
Study 1				
Novices	2.89 (2.47)	.07 (.07)	2.67 (1.58)	1.78 (.83)
Advanced	.60 (.70)	-.02 (.08)	3.30 (.95)	2.60 (1.27)
Study 2				
Cueing	3.09 (1.30)	.11 (.13)	3.27 (.79)	3.55 (.69)
No Cueing	4.80 (2.20)	.29 (.15)	3.50 (.71)	3.90 (.32)

Verbal and Structural Assessment. Study 1: Regarding the verbal assessment both novices and advanced learners perform better on the posttest compared to the pretest (novices: $t(8) = -3.51$, $p <. 01$; advanced: $t(9) = -2.71$, $p < .05$). However, structural assessment shows that after the game only

novices' knowledge structure become more similar to those of experts ($t(8) = -2.77$, $p < .05$). For advanced learners there is no change in similarity ($t(8) = .82$, $p > .05$).

Study 2: Regarding verbal assessment both cueing and no cueing perform better on the posttest compared to the pretest (cueing: $t(10) = -7.88$, $p < .001$; no cueing: $t(9) = -6.90$, $p < .001$). This pattern is also found in structural assessment as cueing and no cueing show an increase in similarity with the knowledge structures of experts (cueing: $t(10) = -2.91$, $p < .05$; no cueing: $t(9) = -5.97$, $p < .001$). However, the increase in terms of effect size for structural assessment seems larger: structural assessment: $d = 1.28$, verbal assessment: $d = .94$.

Verbal and Visual Assessment. Game performance (the game score itself) was positively correlated with performance on the visual questions (Study 1: $r = .61$, $p < .01$; Study 2: $r = .71$, $p < .001$), but not with performance on the four verbal questions (Study 1: $r = .44$, $p > .05$; Study 2: $r = .42$, $p > .05$).

Conclusion and Discussion

Given the small number of participants, the results only provide indications and conclusions can only be drawn with caution. Structural assessment and verbal assessment partly display different patterns in learning results. Study 1 showed an increase in similarity with expert's knowledge structures for novices, but not for advanced learners. This pattern was not reflected in the verbal-oriented performance. In study 2 playing the game yielded an increase in both structural and verbal assessment, but the effect size of the increase in structural assessment was larger. It is difficult to conclude whether structural assessment is more suitable to measure implicit learning. The results suggest that structural assessment measures an individual's understanding of a domain *at least* differently from verbal assessment. The next step involves more research, with more participants and comparisons with post-training performance (e.g., emergency simulations) to further uncover these differences.

There is some evidence for the second hypothesis. The significant correlations between game performance and performance on visual, context-rich items in both studies suggest that performance in the visual game world is positively associated with test items that closely resemble the visual game world. However, the low number of items and the fact that our visual items still involved some verbal processing justify more research with more (mono-dimensional) items. Summarized, the results suggest that structural assessment and the use of visual items are worthwhile in the context of serious games, but more research is needed.

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).

References

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

Spek, E.D. van der, Wouters, P., & Oostendorp, H. van (in press). Code Red: Triage or COgnition-based DEsign Rules Enhancing Decionmaking Training In A Game Environment. *British Journal of Educational Technology*.

Wouters, P., Spek, E.D. van der & Oostendorp, H. van (2009). Current practices in serious game research: A review from a learning outcomes perspective. In Connolly, T.M., Stansfield, M. & Boyle, L. (Eds.), *Games-Based Learning Advancements for Multisensory Human Computer Interfaces: Techniques and Effective Practices* (pp. 232-255). Hershey, PA: IGI Global.