Combining Software Games with Education: Evaluation of its Educational Effectiveness

Maria Virvou, George Katsionis and Konstantinos Manos
Department of Informatics
University of Piraeus, Piraeus 18534, Greece
mvirvou@unipi.gr
gkatsion@kman.gr
konstantinos@kman.gr

Abstract

Computer games are very popular among children and adolescents. In this respect, they could be exploited by educational software designers to render educational software more attractive and motivating. However, it remains to be explored what the educational scope of educational software games is. In this paper, we explore several issues concerning the educational effectiveness, appeal and scope of educational software games through an evaluation study of an Intelligent Tutoring System (ITS) that operates as a virtual reality educational game. The results of the evaluation show that educational virtual reality games can be very motivating while retaining or even improving the educational effects on students. Moreover, one important finding of the study was that the educational effectiveness of the game was particularly high for students who used to have poor performance in the domain taught prior to their learning experience with the game.

Keywords

Educational software games, Virtual reality, Evaluation, Intelligent tutoring systems, Student model

Introduction

The process of learning is a very complex cognitive task that can be very imposing on students since it requires a lot of effort from them. Consequently, they need a lot of motivation to cope with it. In view of this, it is within the benefit of education to create educational software that is interesting and stimulating for students. On the other hand, there is a fast growing area of computer technology, that of computer games, that is extremely appealing to children and adolescents. Indeed, anyone who interacts with children and adolescents in every-day life can easily observe that they like computer games. This is also a view that has been supported by many researchers who have conducted empirical studies (e.g. Mumtaz 2001). Thus the computer games technology could be used to render educational software more motivating and engaging. In this respect, the difficult process of learning could become more amusing.

Indeed, there are many researchers and educators that advocate the use of software games for the purposes of education. Papert (1993) notes that software games teach children that some forms of learning are fast-paced, immensely compelling and rewarding whereas by comparison school strikes many young people as slow and boring. Boyle (1997) points out that games can produce engagement and delight in learning; they thus offer a powerful format for educational environments. Moreover, there are studies that have shown that the use of carefully selected computer games may improve thinking (Aliya 2002). As a result, many researchers have developed games for educational purposes (e.g. Conati & Zhou 2002).

However, the attempts to create educational games have not reached schools yet. There are several reasons for this. At first, not all educators and parents are convinced that educational games can be beneficial to students. Second there are criticisms about the quality of the existing educational games. For example, Brody (1993) points out that the marriage of education and game-like entertainment has produced some not-very-educational games and some not very-entertaining learning activities.

Given the motivational advantages of software games as well as the criticisms that have been made on educational games, there has to be further investigation on the advantages and limitations of software games for education. Such investigation may lead to useful guidelines for the design of effective educational software games. Indeed, educational software games should be designed in such a way that they are educationally beneficial for all students, even those that are not familiar with computer games.

In view of the above, we have conducted an evaluation study on a virtual reality educational game that we have developed, which has been briefly described in (Virvou et al. 2002). The game is called VR-ENGAGE and

teaches students geography. VR-ENGAGE aims at increasing students' engagement by providing a popular and motivating virtual reality environment. In this way, it aims at being more effective in teaching students than other educational software and traditional media of education. The main focus of the research described in this paper is to measure the educational effectiveness of an educational VR-game as compared to educational software that does not incorporate the gaming aspect. The main aim of this comparison is to find out whether the gaming environment may improve education.

The virtual reality educational game

VR-ENGAGE is an Intelligent Tutoring System (ITS) that operates through a virtual reality game. In common with most ITSs it has the main components of an ITS. It has been widely agreed that an ITS, should consist of four components, namely the domain knowledge, the student modelling component, the tutoring component and the user interface (Self, 1999; Wenger, 1987). In the case of VR-ENGAGE, the student modelling component models the student's knowledge and his/her ability to reason plausibly about knowledge acquired on the domain of geography. In this way, while playing, students may practice both their factual knowledge on geography and their reasoning ability and thus they are led to "enjoyable" mastering of knowledge. Domain knowledge is represented in the form of hierarchies that capture the relations between domain concepts. The tutoring component generates advice tailored to the needs of individual students. Finally, the user interface consists of the virtual reality game environment and its gaming features.

The environment of VR-ENGAGE is similar to that of the popular game called "DOOM" (ID-Software 1993) which has many virtual theme worlds with castles and dragons that the player has to navigate through and achieve the goal of reaching the exit. VR-ENGAGE has also many virtual worlds where the student has to navigate through. There are mediaeval castles in foreign lands, castles under the water, corridors and passages through the fire, temples hiding secrets, dungeons and dragons. The main similarity of VR-ENGAGE with computer games like DOOM lies in their use of a 3D-engine. However, VR-ENGAGE unlike DOOM and other computer games of this kind is not violent at all and is connected to an educational application.

VR-ENGAGE communicates its messages to students through animated agents that use speech synthesisers or through windows that display text. When a student is asked a question s/he may type the answer in a dialogue box. The user interface employs two types of animated agent, the dragon, which is the virtual opponent of the player and the angel, which is the virtual companion of the player. Both types of animated agent use synthesised voice as well as written messages. However, their voices are different so that the player may distinguish between them.

The story of VR-ENGAGE incorporates a lot of elements from adventure games. The ultimate goal of a player is to navigate through a virtual world and find the missing pages of the book of wisdom, which is hidden. To achieve the ultimate goal, the player has to be able to go through all the passages of the virtual world that are guarded by dragons and to obtain a score of points, which is higher than a predefined threshold. The total score is the sum of the points that the player has obtained by answering questions.

In particular, while the player is navigating through the virtual world, s/he finds closed doors, which are guarded by dragons as illustrated in the example of Figure 1. A guard dragon poses a question to the player from the domain of geography. If players give a correct answer then they receive full points for this question and the dragon allows them to continue their way through the door, which leads them closer to the "book of wisdom".

However, if a player is not certain about the correct answer, s/he is allowed to ask the dragon for a "negotiation". The student modelling capabilities needed for the negotiation mode of the game are based on a cognitive theory, called "Human Plausible Reasoning theory" (Collins & Michalski 1989). This theory formalises the plausible inferences based on similarities, dissimilarities, generalisations and specialisations that people often use to make plausible guesses about matters that they know partially. Important inference patterns in the theory are the statement transforms. These inferences may lead to either correct or incorrect guesses; in any case these guesses are plausible. The theory is used to simulate the reasoning of students when they give an erroneous answer. If the student is found to have used a known pattern from the theory for his/her answer then this answer is considered as a "plausible" error. Thus, in the case of negotiation, the student is allowed to make a guess for which s/he has to provide a justification. The amount of points that the student is going to receive in the negotiation mode, depends on how close the student's answer is to the correct answer and/or how plausible the reasoning that s/he has used is. The results of the error diagnosis that the system performs, are communicated to the student through the virtual companion agent that appears in situations where the student needs help.

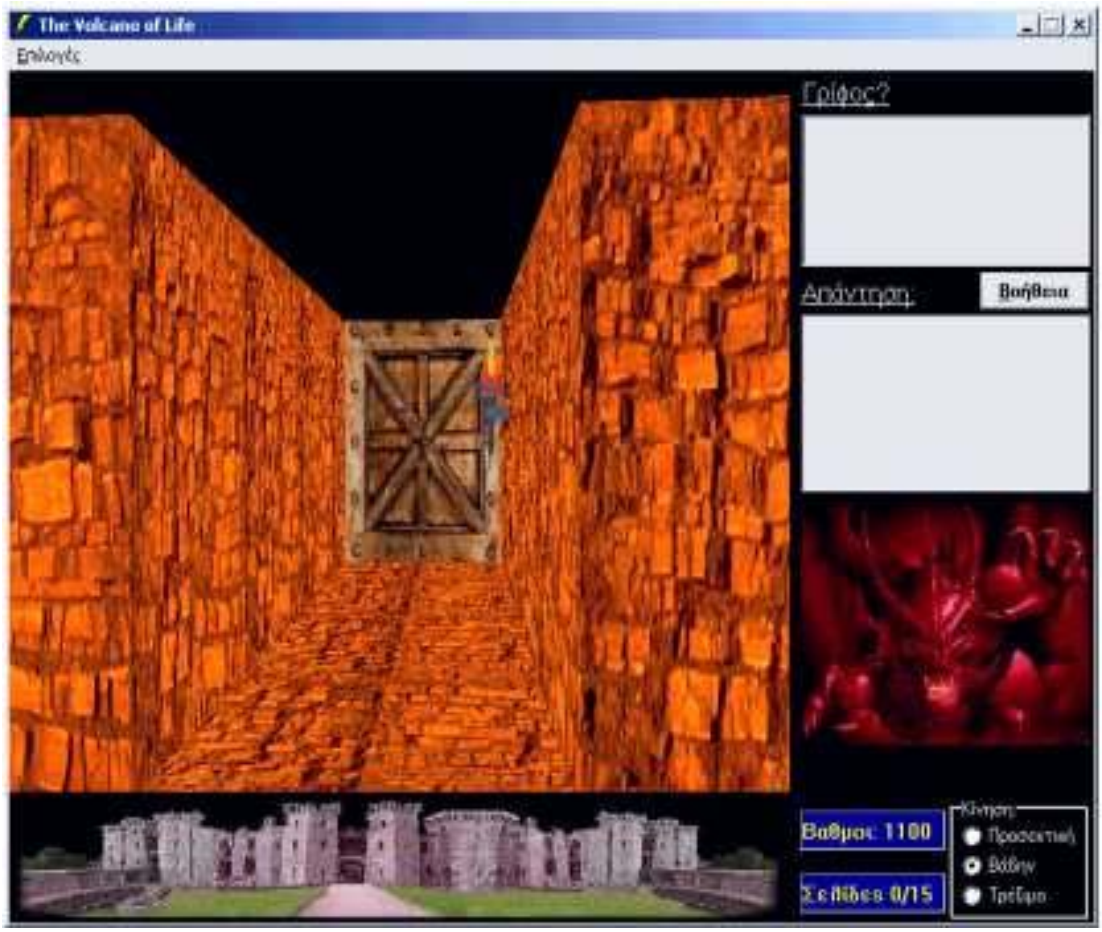

Figure 1. An example of a virtual world of the game

For example, the student may have been asked the following question: "What is the capital town/city of the geographical compartment called Achaia (in Greece)?" While being in the negotiation mode, the student admits that s/he does not know the correct answer and wishes to make a plausible guess such as: "My guess is that *Rio* is the capital of Achaia. I know that Rio is an important town in Achaia. Therefore, it is likely that Rio is the capital of Achaia." This kind of answer is not actually given in natural language and the student is not allowed unlimited selection of relevant pieces knowledge. The student is allowed to select several patterns of reasoning and fill-in the names of cities, towns, mountains, etc. from a list of the domain knowledge of VR-ENGAGE. In this way, it is ensured that the student's answer will be within the "limits" of the domain knowledge encoded in the ITS-game.

For the example above the student would be able to select the city of "Rio" from a list of known Greek cities. The student also provides a justification for his/her guess such as declaring that Rio is an important city of Achaia, and it could be its capital. The student's guess may be correct or incorrect; in the case of the example, it is incorrect because *Patras* is the correct answer. However, the reasoning that s/he has used may reveal whether the student has a good knowledge of geography and whether s/he is able to use it plausibly. In the case of the example, the student has shown that s/he knows some relevant pieces of knowledge, such as that Rio is an important town of Achaia, which is very close to Patras. In this respect, the above answer exhibits some plausible reasoning as compared to an answer that would be absolutely irrelevant. This kind of interaction helps the students to reason about the domain being taught because even if they do not know the correct answer they can use their knowledge on the domain to make a plausible guess about the correct answer.

In this sense the game provides an environment where there is opportunity for a negotiating teaching-learning dialogue between the ITS and the students. Collaborative discourse is an issue that has attracted a lot of research energy in the recent years (e.g. Moore 2000; Baker 1994). The process of becoming an expert in a certain domain should no longer be solely viewed as the acquisition of a representation of correct knowledge; the knowledge to be acquired should flexibly manage open problems (Andriessen & Sandberg 1999). In the case of VR-ENGAGE, the reasoning in the diagnostic process of the system is part of the game's plot and interactivity.

It aims at immersing the student in a way that is educationally beneficial through the teaching learning dialogue between the actors of the game (student and dragon).

As part of the adventure of the game the player may also come across certain objects or animated agents. These objects or animated agents appear at random and give hints to students or guide them to tutoring places respectively. In tutoring places, students are encouraged to read a new part of the domain being taught. However, these hints or the parts of the theory read, are not immediately usable by the students since they refer to questions that the students will have to answer at a location of the virtual world other than the one they are currently at. Hence, the students will have to remember these hints or parts of the theory so that they may use them when the time comes. Educationally, these objects or animated agents motivate students to read and memorise important parts of the theory.

Evaluation aims and experiment

The evaluation that was conducted on VR-ENGAGE focused primarily on evaluating the educational effectiveness of the gaming aspect of the educational software. One could argue that the greatest advantage of games is the motivation provided to students by the game environment whereas one possible disadvantage for the learning process could be the students' distraction by this game environment. However, even the motivational advantage of educational games may be questioned since in a classroom there may be students who do not like games or students who find it difficult to navigate through the virtual world and thus may not be able to benefit to the full from the educational content of the software.

A common theme found in the literature for educational games, both electronic and non-electronic, is that these games and software are considered successful only if they are as effective as traditional classroom education (Mc Grenere 1996). However, this kind of comparison implies that games are not meant to be included in traditional classroom education but rather they are meant to replace it. In our view, games should be used to supplement traditional classroom education. Human teachers still have more abilities in explaining domain issues and diagnosing students' problems than any kind of software irrespective of its sophistication. This view is reinforced by empirical studies that show that no matter how successful an ITS may be, students still prefer the human teacher (e.g. Tsiriga & Virvou 2004). Therefore, in the present evaluation, we did not consider conducting a comparison between human teaching and tutoring through the game.

Thus, to find out whether the game environment is in fact motivating and educationally beneficial to students and not distractive we conducted an experiment where the game-ITS could be compared to an ITS that had a conventional user interface without any virtual reality game. Both educational software applications had the same underlying reasoning mechanisms with respect to student modelling as well as the same help and theory functionalities. The main difference between the two educational software applications (game-ITS and ITS with a simple user interface) was that one had a gaming approach whereas the other one did not have any gaming approach at all. In fact, the software with the simple user interface had a hypertext display of domain theory and exercises that were communicated to students through forms, dialogue boxes, buttons, drop-down menus etc. However, these exercises were not part of any story as in the gaming approach. Moreover, there was no virtual reality environment and no animated-speaking agents. For example, the way that the exam question, "Ethiopia is in Africa. Right or Wrong?", is presented to the user of VR-ENGAGE is illustrated in Figure 2, and the way the same question is presented to the user of the software with a simple user interface is illustrated in Figure 3.

The evaluation experiment was connected to the underlying rationale of the educational game, which was to engage students in learning the domain concepts that were taught to them. Thus the aim of the experiment was to find out whether the educational game was in fact more motivating while it was at least as effective with respect to students' learning as the educational software with the conventional interface. Moreover, one of the primary aims of the experiment was to reveal the degree of educational effectiveness (if any) for students whose performance was considered good, mediocre or poor respectively by their human teachers.

This experiment took place in classrooms. School children usually have a preconception of educational means as being totally different from entertainment. In this respect, the experiment aimed at finding out how school children would react to an educational game in the settings of a real classroom where an entertaining aspect of education would be rather unexpected. This was the main reason why the experiment took place in school-classrooms. Human tutors were present and they were asked to observe their students while they interacted with the computer but were not actively involved in the evaluation. There were, however, lab assistants that helped students with the interaction with the game if the students needed help.

Figure 2. A question posed in the VR-ENGAGE

Figure 3. A question posed in the simple UI application

The experiment consisted of four parts. All four parts were similarly set up and involved a comparison between VR-ENGAGE and the ITS with the simple User Interface (UI) in terms of the educational effectiveness and motivation. All four parts were conducted in parallel. All of the children who participated in all four parts of the experiment were of 9-10 years old and attended the fourth grade of elementary schools in Greece. They had been taught the same syllabus on geography and they had a similar background on the use of computers. More specifically, all of them were computer-literate and had been trained in their respective schools in the use of Windows, the Internet and other popular software packages such as word-processors etc.

Each part of the experiment was different from the other parts in the type of the school-children that participated in it. Specifically, the first part of the experiment involved all the students of 5 classes of school children of the fourth grade of an elementary school, 90 children altogether, and their respective geography teachers.

For the second, third and fourth part of the experiment, the students who participated, were also of 9-10 years old (fourth grade of the elementary school), but the selection of them was based on the mark that these students had received by their respective human teachers in geography in the previous term. The term-marks that the students of the fourth grade of elementary schools receive usually range from A to C. "A" is given to students with good performance, "B" is given to students with mediocre performance and "C" is given to students with poor performance. The participants were selected from the total students of 7 classes (127 students) of the fourth grade of an elementary school, which was different from the one that had been used for the first part of the experiment. From the total of the 127 students, 30 students were selected for the second part of the experiment based on the criterion of their having received the mark "A" in the previous term, 30 students were selected for the third part of the experiment based on the criterion of their having received the mark "B" in the previous term and finally 30 students were selected for the third part of the experiment based on the criterion of their having received the mark "C" in the previous term. The number 30 was selected so that we could have equal numbers of students participating in each of the remaining three parts of the experiment.

Each group of children that were selected to participate in each part of the experiment was randomly divided into two independent sub-groups of the same number of children. Thus, there were two independent sub-groups of 45 students for the first part of the experiment, two independent sub-groups of 15 students for the second part of the experiment, two independent sub-groups of 15 students for the third part of the experiment and two independent sub-groups of 15 students for the fourth part of the experiment. The first sub-group of each group would use VR-ENGAGE and the second sub-group of each group would use the ITS with the simple UI (User Interface).

Before using their respective version of educational software, the students of both sub-groups of all the groups, were asked to work on a pre-test using paper and pencil. This pre-test was an ordinary classroom test in which every student had to answer 100 questions by filling in a test paper. The students' performance in the pre-test was compared to the students' performance in a post-test that was given to the students after the use of their respective software. The post-test was of a similar level of difficulty as the pre-test and consisted of the same number of questions (100). The comparison of students' results in the pre-test and the post-test was used to draw conclusions about the educational effectiveness of VR-ENGAGE as compared to the simple ITS. In particular, the school teachers were asked to count the number of erroneous answers of each student in the pre-test and the post-test.

The students' pre-test and post-test performance was compared using t-test statistics. In particular, the educational effect of VR-ENGAGE was compared to that of the simple ITS by comparing the number of mistakes of the students of the VR-ENGAGE sub-groups with the number of mistakes of the students of the respective sub-groups that had used the simple ITS. It was expected that the number of mistakes that the students would make after the use of either of the software versions would be reduced in comparison with the pre-test because both applications provided quite sophisticated tutoring from the adaptive presentation of the theory and the reasoning and student modelling of both applications. However, the post-test would reveal the degree to which students who had used VR-ENGAGE exhibited greater or less improvement than those who had used the ITS with the standardised user interface.

After the post-test, all the students who had participated in the experiment were also interviewed about their experiences using their respective educational software. Moreover, the teachers of the school classes who had participated in the experiment were also interviewed concerning their students' performance and behaviour during the experiment. Teachers were also asked to give their comments on their students' performance on the pre-tests and post-tests.

Evaluation results

First part of the evaluation

As mentioned earlier, the first part of the evaluation involved 90 students of the fourth grade of an elementary school who were separated into two sub-groups of 45 children that would use the VR-ENGAGE and the simple ITS respectively. The results showed a greater improvement of the VR-ENGAGE users over the users of the other software. In particular, in the post-test, the players of VR-ENGAGE made 43.15% less mistakes than in the pre-test. The other sub-group of students that had used the simple ITS resulted in 32.48% less mistakes of their answers in total, as compared to the pre-test. Thus the players of VR-ENGAGE resulted in a higher improvement of 10.67% in terms of their mistakes than the users of the simple ITS. This showed that VR-ENGAGE had achieved its aim of being at least as effective as conventional educational software in the learning outcomes and was in fact better in this respect.

In more detail, the total questions that were asked to the total number of students of each sub-group were 4500: *45 students x 100 questions = 4500 questions.* In total, the students who had worked with VR-ENGAGE failed during the pre-test in *1599* questions. The mean value of errors per student was 35.53 and the standard deviation 18.51. In total, the students who had worked with the simple ITS failed during the pre-test in *1647* questions. The mean value of errors per student was 36.6 and the standard deviation 19.23. An initial analysis concerning the comparison of the number of mistakes of each sub-group in the pre-test was not statistically significant showing that the two sub-groups had similar background knowledge on geography. Indeed, there was a t-test performed for the pre-test of the students of the two sub-groups. The null hypothesis (H_0) was that there was no difference between the mistakes of the two sub-groups and the research hypothesis, (H_1) was that there was a difference between the mistakes of the two sub-groups. The t-value result of 0.27 was smaller than its critical value 2.00. This showed that the students of the two sub-groups had similar prior knowledge of the domain of geography.

Then, after the students had completed their interactions with the two applications, they were given the post-test. The players of VR-ENGAGE made 909 mistakes in the post-test. This number of mistakes as compared to the 1599 of the pre-test constituted an improvement of 43.15%. The students who had worked with the other educational software made 1112 mistakes, which constituted an improvement of 32.48% in the number of erroneous answers. The users of both educational applications showed an improvement in their post-test performance. This was expected since both applications provided quite sophisticated tutoring and reasoning mechanisms. However, the main aim of our experiment was to compare the improvement of the respective sub-groups that had used the two applications.

Thus, the second statistical analysis compared the improvement between VR-ENGAGE and the simple ITS users, on the number of mistakes for each sub-group between the pre-test and the post-test. The comparison concerned the improvement that the users of VR-ENGAGE had vs the improvement that the users of the other educational application had. There was a t-test performed. The null hypothesis (H_0) was that there was no difference in the improvement on the number of mistakes for the two sub-groups (VR-ENGAGE and simple ITS) and the research hypothesis, (H_1) was that there was a difference in the improvement on the number of mistakes for the two sub-groups. As a result, the t-value of 4.52 was significantly greater than its critical value 2.00. This showed that the difference was statistically significant for the first sub-group in comparison with the difference of the second sub-group, leading to the result that the 45 students who had used VR-ENGAGE had a higher educational benefit than the 45 students who had used the simple ITS. All the statistical results are summarised in Table 1.

Specifically, Table 1 illustrates the mean value of the errors made during the pre-test, the mean value of the errors made during the post-test and the mean value of the percentage improvement on mistakes between the two tests, for both of the sub-groups, the first who had used VR-ENGAGE and the other who had used the simple ITS. Additionally it includes the respective results of the t-tests. These are the results of the first t-test, after the pre-test, which show that there was no significant difference on the background knowledge on geography for the two sub-groups of VR-ENGAGE and the simple ITS, and the results of the second t-test, after the post-test, which show that there was a greater improvement on the number of mistakes for the sub-group of VR-ENGAGE users over the other group. These results involve the standard deviations, the t- values (Tv) and the critical values (Cv) of the t-tests.

Table 1. Results of the analysis of the students' mistakes

Variable	VR-ENGAGE Sub-Group (n=45)		Simple ITS Sub-Group (n=45)		Tv; Cv
	Mean Value	*Standard Deviation*	*Mean Value*	*Standard Deviation*	
Pre-test errors (Between 0 and 100)	35.53	18.51	36.60	19.23	Tv = 0.27; Cv = 2.00
Post-test errors (Between 0 and 100)	20.20	10.21	24.71	14.07	
Improvement percentage on mistakes between pre-test and post-test	43.15%	12.57	32.48%	9.26	Tv = 4.52; Cv = 2.00

In the above t-tests the t-value of each t-test is calculated by performing a t-test for independent samples for each of the null and research hypotheses (Voelker, 2001). The critical value for each t-test is the value taken from Table T for a two-tailed research hypothesis depending on the sample number. The t-test results show that there is a statistically significant difference in favour of the educational benefits of VR-ENGAGE over the simple ITS.

Second, third and fourth part of the evaluation

The second, third and fourth part of the evaluation involved 90 students of the fourth grade of an elementary school, which was different from the first part of the evaluation, separated into three groups of 30 children having poor, mediocre and good performance in geography. Every group of 30 children was then separated into two sub-groups of 15 children that would use the VR-ENGAGE and the simple ITS respectively.

In the post-test, the VR-ENGAGE students who used to be poor and mediocre performers made 48.97% and 38.5% less mistakes respectively than in the pre-test. The students of poor and mediocre academic performance that had used the simple ITS resulted in 31.57% and 31.64% less mistakes respectively, as compared to the pre-test. Thus the sub-groups of students, of previous poor and mediocre academic performance that had used VR-ENGAGE resulted in a higher improvement of 17.4% and 6.86% respectively in terms of their mistakes than the students of the respective sub-groups that had used the other application. Moreover, the good students who had used VR-ENGAGE resulted in a 33.8% improvement while the good students who had used the other application resulted in a 32.84% improvement. This showed, that for the two sub-groups of the good students there was also a small difference in favour of VR-ENGAGE but this difference was not statistically significant.

In particular, the total questions that were asked to the total number of students of each sub-group were 1500: *15 students x 100 questions = 1500 questions.* In total, the sub-groups of students of previous poor, mediocre and good academic performance who had worked with VR-ENGAGE failed during the pre-test in *921, 535 and 213* questions respectively. In total, the sub-groups of students of previous poor, mediocre and good academic performance who worked with the conventional educational software failed during the pre-test in *906, 493 and 201* questions respectively. An initial analysis concerning the number of mistakes of each sub-group in the pre-test involved 3 t-tests for the students of poor, mediocre and good academic performance respectively. For each of the 3 t-tests the null hypothesis (H_0) was that there was no difference between the mistakes of the VR-ENGAGE sub-group and the simple ITS sub-group. The research hypothesis, (H_1) was that there was a difference between the mistakes of the two sub-groups. The t-value results of 0.53 for the students of poor previous performance, 1.27 for the students of mediocre previous performance and 0.56 for the good students were smaller than their critical values of 2.05, 2.05 and 2.05 respectively. This led to the acceptance of the null hypothesis (H_0), which showed that the respective sub-groups, that had used VR-ENGAGE and the simple ITS, for each of the three categories of students, had similar prior knowledge of the domain of geography.

Then, after the students had completed their interactions with the two applications and answered the questions of the post-test, we came up with the following results. The players of VR-ENGAGE (poor, mediocre and good previous performance) made 470, 329 and 141 mistakes respectively. These mistakes compared to the 921, 535 and 213 in the pre-test constituted an improvement of 48.97%, 38.5%, and 33.8% respectively. The students who had worked with the non-game ITS made 620, 337 and 135 mistakes, which constituted an improvement of 31.57%, 31.64% and 32.84% respectively in the number of answers failed. The statistical analysis which took

place after the post-test, compared the improvement on the number of mistakes for each sub-group between the pre-test and the post-test. The comparison concerned the improvement of the users of each of the three categories of students of VR-ENGAGE vs the improvement of the users of each of the three categories of students of the other educational application.

There were 3 t-tests performed concerning the comparison of the improvement on the number of mistakes for the students of poor, mediocre and good academic performance respectively. For each of the 3 t-tests the null hypothesis (H_0) was that there was no difference in the improvement on the number of mistakes for the respective sub-groups. The research hypothesis, (H_1) was that there was a difference in the improvement on the number of mistakes for the respective sub-groups. The t-value result of 4.86 for the poor performing students was significantly greater than its critical value of 2.05. The t-value result of 2.28 for the average students was adequately greater than its critical value of 2.05. This showed that the difference in the improvement on the number of mistakes for the respective sub-groups of both poor and mediocre performing students was statistically significant. However, the difference on the improvement was mostly evident for the case of poor performing students. On the other hand, the t-value result of 0.27 for the good students was significantly smaller than its critical value of 2.05. This showed that the difference in the improvement on the number of mistakes, was not statistically significant for the respective sub-groups of good students, leading to the result that good students who had used VR-ENGAGE benefited in a similar way with the good students that had used the non-game ITS.

The mean values of the pre-tests and the post-tests of the students, and the results of the above t-tests are summarised in Table 2. In particular, Table 2 illustrates the mean values of the errors made during the pre-tests, the mean values of the errors made during the post-tests and the mean value of the percentage improvement on mistakes between the test pairs, for all of the sub-groups pairs, the first who had used VR-ENGAGE and the other who had used the simple UI application. Additionally it includes the respective results of the t-tests. These are the results of the three t-tests, after the pre-tests, which showed that there was no significant difference on the background knowledge on geography for the sub-group pairs, and the results of the three t-tests, after the post-tests, which showed the results of the comparison between the improvement on the number of mistakes for VR-ENGAGE users and the users of the simple ITS. These results involve the standard deviations, the t-values (Tv) and the critical values (Cv) of the t-tests.

Table 2. Results of the analysis of the students' mistakes

Variable	**VR-ENGAGE Sub-Group (n=45)**		**Simple ITS Sub-Group (n=45)**		**Tv; Cv**
	Mean Value	*Standard Deviation*	*Mean Value*	*Standard Deviation*	
Pre-test errors of students of previously **poor academic performance** (Between 0 and 100)	61.40	7.70	60.40	5.84	Tv = 0.53; Cv = 2.05
Pre-test errors of students of previously **mediocre academic performance** (Between 0 and 100)	35.67	9.43	32.87	8.62	Tv = 1.27; Cv = 2.05
Pre-test errors of students of previously **good academic performance** (Between 0 and 100)	14.20	4.36	13.40	3.40	Tv = 0.56; Cv = 2.05
Post-test errors of students of previously **poor academic performance** (Between 0 and 100)	31.33	8.04	41.33	6.00	
Post-test errors of students of previously **mediocre academic performance** (Between 0 and 100)	21.93	8.18	22.47	6.93	
Post-test errors of students of previously **good academic**	9.40	3.02	9.00	1.93	

performance (Between 0 and 100)					
Improvement percentage on mistakes between pre-test and post-test for students of previously **poor academic performance**	48.97%	10.94	31.57%	7.72	Tv = 4.86; Cv = 2.05
Improvement percentage on mistakes between pre-test and post-test for students of previously **mediocre academic performance**	38.50%	10.06	31.64%	5.08	Tv = 2.28; Cv = 2.05
Improvement percentage on mistakes between pre-test and post-test for students of previously **good academic performance**	33.80%	9.66	32.84%	9.67	Tv = 0.27; Cv = 2.05

In the above t-tests the t-value of each t-test is calculated by performing a t-test for independent samples for each of the null and research hypotheses. The samples were independent (and not correlated) because the experiment aimed at comparing the improvement of two independent groups of students, the group of VR-ENGAGE users and the group of the simple ITS users. The critical value for each t-test is the value taken from Table T for a two-tailed research hypothesis depending on the sample number.

In summary, the above results showed that the sub-group of students of previous poor performance, which had used VR-ENGAGE, benefited the most of all the sub-groups from the educational game. In addition, the sub-group of VR-ENGAGE students of previous average performance had also benefited more than the respective sub-group that had used the simple ITS since they made fewer mistakes (and the difference was statistically significant). On the other hand, good students who had used VR-ENGAGE benefited in a similar way with the good students that had used the non-game ITS.

Interviews of students and teachers

All of the students who had participated in the experiment were interviewed concerning the software they had used. These interviews revealed that the players of VR-ENGAGE were fascinated by the idea of a game in the classroom and they were certainly more enthusiastic about the software that they had used than the other group of students. However, despite the fact that all students had liked the game in the context of their classroom work, a large part of them criticised the game in comparison with other commercial games and said that they would like VR-ENGAGE to have more virtual objects, a more sophisticated environment, more adventure and more action. Students who were experienced game-players mainly made these comments. Such students had high expectations from VR-games.

As to the teachers, most of them were particularly impressed by the effect that the game had on students who were previously poor performers on geography. This category of students included quite a lot of those students who the teachers thought were not easily disciplined in class. The teachers reported that these students seemed absolutely absorbed by the game environment and kept working in piece and quiet without talking to anyone and without disturbing anyone. To some extent, this comment was also made for the same category of students who were given the other educational application to work with. In general, the teachers thought that the use of computers was very good for the students who they used to consider as non-disciplined in class. However, they thought that those who had used the game seemed so immersed that their behaviour in class had changed completely and they had appeared to be very satisfied and interested in the educational content. The teachers were very happy with their students' performance on the post-test and most of them said that they would certainly wish to include educational games of this kind in classroom. Some of them suggested that they might even use the game on their own laptop in classroom and show the action of the game through a projector to their whole class so that the whole class could participate in a single game play.

Discussion and conclusions

The results from the evaluation showed that students would benefit from educational games in classrooms and would be quite happy to work with a computer game, which represents a more amusing teaching fashion than that of conventional educational software. Moreover, one important finding that should be noted from the t-tests of the second, third and fourth part of the evaluation is that when the subgroups of students who previously had good, average and poor performance respectively were compared separately, it was revealed that the subgroup of students who used to be poor performers had benefited the most from the game environment whereas the subgroup of good students had benefited the least from the game environment. This coincides with findings about the benefits of multimedia in general (Mayer, 2001).

The above finding may be explained by the fact that good students usually perform well under any circumstances, whereas the rest of the students and particularly those who perform poorly may do so because of lack of interest for their lessons and tests. Thus, students with little interest in their courses may benefit from extra motivating environments such as those of VR-educational games. This finding was also confirmed by the teachers' impression about the students who they thought they were not easily disciplined in class. These students were reported to have been absorbed by the game and they did not seem willing to take time out to talk to other students or to try to cheat on the test and so on. This is probably due to the fact that games are able to attract the attention of students who do not concentrate easily on their assignments due to boredom or other distractions.

The students who used to have good academic performance did not have any significant difference in their improvement through the use of the game or the use of the other software. However, one important finding is that the performance of previously good students has not deteriorated by the use of the educational game due to possible usability problems in the VR-environment or their possible distraction through the game. It seems that good academic performers can keep their good academic record despite the fact that some of them were not experienced virtual reality game players. From the interviews it was evident that they too had enjoyed the learning experience through the game to a large extent.

However, it must be noted that during the experiment all students had as much help as they needed from lab instructors concerning their interaction with the VR environment of the game. If the students had used the software on their own at home, then perhaps they might have had more usability problems, especially those who were not sufficiently experienced in virtual reality game playing. Such problems might have resulted in less good educational results. Therefore, in future versions of VR-ENGAGE we aim to improve the usability of the game environment and incorporate more on-line help. Finally, the game environment of the educational game has to be very competitive with commercial games to attract a high degree of interest from students. This is so because children are quite familiar with commercial games and therefore they have high expectations from game environments.

References

Aliya, S. K. (2002). The role of computer games in the development of theoretical analysis, flexibility and reflective thinking in children: A longitudinal study. *International Journal of Psychophysiology, 45*, 149.

Andriessen J., & Sandberg, J. (1999). Where is Education Heading and How about AI? *International Journal of Artificial Intelligence in Education, 10*, 130-150.

Baker, M. (1994). A Model for Negotiation in Teaching-Learning Dialogues. *International Journal of Artificial Intelligence in Education, 5* (2), 199-254.

Boyle, T. (1997). *Design for Multimedia Learning*, London: Prentice Hall.

Brody, H. (1993). Video Games that Teach? *Technology Review, November/December*, 51-57.

Collins, A. & Michalski, R. (1989). The Logic of Plausible Reasoning: A core Theory. *Cognitive Science, 13*, 1-49.

Conati, C., & Zhou, X. (2002). Modeling students' emotions from cognitive appraisal in educational games. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of the Intelligent Tutoring Systems 2002,* Lecture Notes in Computer Science, 2363, Berlin Heidelberg: Springer, 944-954.

Hasebrook, J. P., & Gremm, M. (1999). Multimedia for Vocational Guidance: Effects of Indiviualised Testing, Videos and Photography on Acceptance and Recall. *Journal of Educational Multimedia and Hypermedia, 8* (4), 377-400.

Id Software (1993). *Doom - Virtual Reality Computer Game*, Id Software Company, Texas, USA.

Mc Grenere, J. L. (1996). Design: Educational Electronic Multi-Player Games. A Literature Review *Technical Report 96-12*, Vancouver, Canada, University of British Columbia.

Mayer, R. E. (2001). *Multimedia Learning*, New York: Cambridge University Press.

Moore, D. (2000). A framework for using multimedia within argumentation systems. *Journal of Educational Multimedia and Hypermedia, 9* (2), 83-98.

Mumtaz, S. (2001). Children's enjoyment and perception of computer use in the home and the school. *Computers and Education, 36*, 347-362.

Papert, S. (1993). *The Children's Machine: Rethinking School in the Age of the Computers*, New York: Basic Books.

Tsiriga, V. & Virvou, M. (2004). Evaluating the intelligent features of a web-based intelligent computer assisted language learning. *International Journal of Artificial Intelligence Tools, 13* (2), 411-425.

Virvou, M., Manos, C., Katsionis, G., & Tourtoglou, K. (2002). VR-ENGAGE: A Virtual Reality Educational Game that Incorporates Intelligence. *Paper presented at the IEEE International Conference on Advanced Learning Technologies (2002)*, September 16-19, 2002, Kazan, Russia.

Voelker, D. (2001). *Statistics*, New York: Wiley.

Wenger, E. (1987). *Artificial Intelligence and Tutoring Systems*, Los Altos, CA, USA: Morgan Kaufmann.