Learning vocabulary through a serious game in Primary Education

Maaike Heitink, Petra Fisser and Joke Voogt
University of Twente
Netherlands
m.c.heitink@utwente.nl
p.h.g.fisser@utwente.nl
j.m.voogt@utwente.nl

Abstract: This study explored the effect of a serious game on the vocabulary of K4-6 students in primary education. 206 students and 10 teachers used the serious game 'Word Score 2' during vocabulary lessons in three different conditions: (a) online game and vocabulary instruction, (b) online game only, and (c) paper game and vocabulary instruction. In every condition students' vocabulary was tested before, directly after and four weeks after the lessons took place. Additionally a student questionnaire and teacher interview regarding their experiences has been employed. Results show a significant learning effect for conditions in which teachers used vocabulary instruction additional to the game (both paper and digital). Comparison between the three conditions showed the highest learning effect on both the post- and retention test was achieved by students that played the online game and received the corresponding instruction. Teachers were excited about students' performance and enthusiasm. All teachers thought Word Score 2 fit their usual program and would be willing to replace the conventional vocabulary method.

Introduction

Various studies show that the usage of serious games in education has a positive effect on students' performance as students are working on challenging problems in an active and motivated way (e.g. Boyan & Sherry, 2011; Leemkuil, 2006). As vocabulary education has impact on many educational subjects and thus plays an important part in students' school performances (Cöp, 2009), a serious game concerning vocabulary education can be very effective.

This study investigates the effect of the serious game 'Word Score 2' on the vocabulary of students in primary education. Word Score 2 is an online serious game with supporting educational materials, intended to increase the vocabulary of students at the age of 9-12 in a playful way. The game and the additional instructional materials are based on the model of Verhallen and Verhallen (1994), that suggest new words should be presented in four stages: (1) activating prior knowledge, (2) explaining new words, (3) consolidating new words and (4) testing whether new words have become part of students' daily vocabulary. In Word Score 2, the first two stages are covered by classroom instruction and the consolidating and testing stages are integrated in the game. Students practice new words while playing different mini games and vocabulary tests take place before and after the start of a new theme. Word score 2 includes 600 words divided in 10 themes, intended for K6. Every theme lasts 4 weeks. This study focused on only one of the themes, namely: Halloween.

Theoretical Underpinnings

Serious gaming

Games are competitive, situated, interactive learning environments based on a set of rules and / or an underlying model, which with certain limitations and uncertain conditions, a challenging goal to be achieved (Leemkuil, 2006). An educational game is a game designed for educational purposes (Van Rooij, Jansz & Schoenmakers, 2010). There are many different games and different ways to classify these games. Kebritchi and

Hirumi (2008) for instance categorize games with regard to the underpinning instructional theory; with direct instruction the player of the game is explained exactly what to do. In experiential learning, the player of the game learns by gaining experience while performing tasks. In discovery learning, the player has to search and learn how the game and the missions are put together by him or herself. In situated learning the game is placed within a certain context and constructivism assumes that the player must construct his or her own knowledge and in these games the player is therefore involved in some levels of game design. Research has shown that the challenge to solve cognitive puzzles is the main motivation for playing games. The challenges in a game can be both visible and hidden. Visible challenges are for example the tasks that the player has to perform. The hidden challenges are discovering the unique rules, opportunities, constraints and strategies of the game. To overcome the visible challenges, the player must first master the hidden challenges and in order to be able to do this the player must create a mental model of the obstacles built into the game world. This way of learning is very active, complex, and learner-centered. In addition, Rieber (in Hirumi et al, 2010) states that a very important element of a game is that if the player overcomes the challenges and enjoys playing it, he or she will make a good effort to play better and more often.

Serious games and vocabulary

Different studies emphasize the importance of vocabulary education as lacking vocabulary often leads to poor learning performance in many other educational subjects (e.g. Cöp, 2009). A multitudinous learning environment with many visible and accessible vocabulary related items is therefore very important. ICT can contribute to these environments and thus eventually to expanding students' vocabulary. McCardle and Chhabra (2004) show that computer based vocabulary instruction can be more effective than direct instruction, dependent a combination of the software's quality and the characteristics of the target audience. Most important factors for effective serious games in vocabulary education is interaction and motivation (Peterson, 2010), where interaction is mostly focusing on the social nature of the game and motivation has the goal to make learning easier and keep students involved.

Research question

This study aims to answer the following research question: "what are the learning effects of the Word Score 2 game on the vocabulary of the students?". To answer this question, both immediate learning and retention effects were examined for every condition and results were compared between conditions. Additionally, because performance is depended on how both the teachers and students experience the game, their experiences are examined as well.

Method

A comparative study was employed to study the effect of the game and corresponding instruction on students' vocabulary. Word Score 2 was implemented in nine classrooms on eight schools. In four weeks the game and corresponding instruction offered students 60 words, intended for K4, matching the theme 'Halloween'. Schools participated in the following three conditions: (A) Schools where teachers used the Word Score 2 instruction and where students played the Word Score 2 game, (B) schools where teachers did not use the Word Score 2 instruction and where students played the Word Score 2 game and (C) schools where teachers used the Word Score 2 instruction and where students trained their vocabulary by playing a paper version of the Word Score 2 game. Respondents for this research were 206 students of 9-12 years old (K4-6) and 10 teachers. An overview is provided in Table 1.

Condition	School	Classes	Teachers	Groups	Students
Digital game + instruction (A)	A1	1	1	K3-4	24
	A2	2	2	K4	44
	A3	1	2	K4	33
Digital game (B)	B1	1	1	K4	17
	B2	1	1	K4	14
	B3	1	1	K4	24
Paper game + instruction (C)	C1	1	1	K4-6	24
	C2	1	1	K4-6	26
Total		9	10		206

Table 1: Demographic data

Most students speak the language of the game (Dutch) in their home situation (always: 66.3%, or almost always: 27.1%). Only a minority of the students (6.6%) never or almost never speak Dutch at home. The distribution between sexes was about equal.

Instruments

Data was collected through vocabulary tests, interviews, and questionnaires to foster triangulation. Before the start of the intervention teachers followed a workshop regarding the technical and pedagogical aspects of the game and students filled out a questionnaire regarding their language situation at home and their computer usage/attitudes. Additionally students' vocabulary level was determined by a vocabulary test. The pre vocabulary test was based on the words and descriptions of Word Score 2 and included 20 multiple choice items with four answer options.

Directly after the intervention, both a teacher interview and a student questionnaire were administered, and another vocabulary tests took place. To avoid any pretest learning effects, the post vocabulary test resembled the pretest, but included 10 words in items with different question forms and 10 new words. Data about teacher experiences were collected through a structured interview. The interview included questions about the level of the game and instruction, usability of the game and instruction, implementation in practice (preparation, instruction an game), integration in their education, perception about students' motivation and performance and ideas about students playing the game at home and involving parents. The students' questionnaire included 14 four-point Likert scale questions regarding their perception on the instructiveness, difficulty and appreciation of Word Score 2. Questions in the questionnaire were adapted to the different conditions (e.g. no questions about the teacher's lessons for students in condition B).

Four weeks after Word Score 2 had finished, a retention vocabulary test took place to investigate to what extent Word Score 2 impacted students vocabulary on a long term. This test matched the post vocabulary test.

Data analysis

Test and item analysis (TIA) was done to determine the quality of the vocabulary tests. The TIA showed that the reliability of the pre- and posttest was good (GLB=0.85, N=40) and average (GLB=0.77, N=151) respectively. Because of technical difficulties, complete results of only 40 participants were available for the pretest, not enough for a reliable TIA. Therefore the results of the pretest's TIA can only be considered as an indication. Although the tests turned out to be relatively easy for the population (pretest: p=0.76, posttest: p=0.83), both pre- and posttests made sufficient (Rit=0.29) and good (Rit = 0.36) distinction between high-scoring students and low-scoring students. Qualifications are corresponding to Van Berkel & Bax (2006).

Means and standard deviations of the scores on the pre-, post- and retention test were calculated to determine learning effects. A paired sampled t-test was used to determine learning effects for every condition separately. Because an ANOVA and posthoc analysis (Bonferroni) showed significant differences for the pretest results between all three conditions (F(2,191)=26.73, p<0.0001), an ANCOVA was performed using the posttest values as dependent variable, the pretest/retention test values as covariate, and 'condition' (with three levels) as independent variable. The data satisfied the assumed model as no evidence was found against the assumption that the regression lines of the dependent variable on the covariate have the same slope, F (2, 146) = 1.1615, p = 0.20. Another, similar, ANCOVA was performed to determine differences in learning effect between sexes.

Effect sizes (Cohen's d (Cohen, 1960)) were calculated for every condition between pre- and posttest, retention- and posttest and retention- and pretest. Because of the pretest differences between the conditions, corrected effect sizes where calculated to determine the differences in effect sizes between the three conditions. Effect sizes d corrected for pretest differences were calculated as were corrected $d_{corr}=d_{posttest} - d_{pretest}$ with d reflected the difference between the mean test scores of two conditions divided by the pooled standard deviation.

Students' experiences are described by means and standard deviations. Teachers' interview results are summarized in a table and described.

Results

Learning results

Vocabulary tests were used to determine students' vocabulary performance. Results for every condition are shown in Table 2 to Table 4. No significant differences in learning effects were found between sexes. Results of

conditions A and C prove significant learning effects and high effect sizes on the posttest compared to the pretest. No significant learning effect could be determined for condition B (see Table 2).

Condition		**Pretest**		**Posttest**		***t***	***p***	***d***
	N	*M*	*SD*	*M*	*SD*			
A	89	12.72	2.996	15.49	3.237	8.47	0.0001	0.89
B	28	9.61	3.695	8.29	4.971	-1.78	0.0870	-0.30
C	35	15.11	1.843	16.97	1.948	5.23	0.0001	0.98

Table 2: Learning effects for every condition, posttest compared to pretest (*note: no* posttest results for B1 due to technical difficulties).

Comparing the results on the retention test to the results on the posttest shows a long term learning effect for conditions A and B, with a medium effect for condition A and a large effect for condition B. This is however not the case for condition C, where no significant differences were found (see Table 3).

Condition		**Posttest**		**Retention test**		***t***	***p***	***d***
	N	*M*	*SD*	*M*	*SD*			
A	81	15.54	3.248	17.11	1.904	4.95	0.0001	0.59
B	11	8.55	4.413	14.55	1.440	4.45	0.001	1.83
C	44	17.27	1.945	16.75	2.373	-1.86	0.069	-0.24

Table 3: Learning effects for every condition, retention test compared to posttest(*note: no* posttest results for B1 and retention test results for B3 due to technical difficulties).

Learning effects between the retention test and pretest confirm this image: results of all conditions show a significant learning effect compared to the pretest and a large effect sizes (see Table 4).

Condition		**Pretest**		**Retention test**		***t***	***p***	***d***
	N	*M*	*SD*	*M*	*SD*			
A	93	12.92	3.037	17.12	1.822	14.16	0.0001	1.67
B	29	11.24	3.471	14.79	2.094	5.53	0.0001	1.24
C	37	15.43	1.894	16.81	1.883	4.30	0.0001	0.73

Table 4: Learning effects for every condition, retention test compared to pretest(*note: no* retention test results for B3 due to technical difficulties).

Differences in learning effects between the three conditions are shown in Table 5. Data proves significant differences between the conditions after correcting the posttest values for pretest results, $F(2, 148) = 31.12$, $p < .0001$. After correcting the retention test for the pretest, significant differences between the three conditions were found as well, $F(2,155) = 14.05$, $p < .0001$.

	Condition A			**Condition B**			**Condition C**		
	Pre (*N*=101)	post (*N*=89)	ret (*N*=93)	pre (*N*=55)	post (*N*=28)	ret (*N*=29)	pre (*N*=38)	post (*N*=35)	ret (*N*=37)
M	12.85	15.49	17.12	10.75	8.29	14.79	15.34	16.97	16.81
SD	3.00	3.24	1.82	3.51	4.97	2.09	1.95	1.95	1.88
M_{corr}		15.48	17.18		10.21	15.28		15.47	16.26

Table 5: Mean scores and standard deviations of vocabulary tests for every condition. M_{corr} is the corrected mean (corrected for the pretest).

Posthoc analysis (Bonferroni) show a significant differences on the posttest between conditions A and B ($p < .0001$) and conditions B and C ($p < 0.0001$). However, no significant differences were found between A and C. On the retention test significant differences were found between conditions A and B ($p < 0.0001$) and conditions A and

C ($p < 0.034$), but no significant differences could be found between conditions B and C. Since the words in the game were intended for K4 students, a separate analysis has been done for the data of K4 students only. Results where similar.

Calculating the effect sizes between the three conditions proved that both directly after (posttest) and four weeks after (retention test) Word Score 2 finished, the data of students in condition A (digital game + instruction) resulted in the highest learning effect. Results of effect size calculations are shown in Table 6, where Cohen's rule of thumb is used for interpretation (Cohen, 1960).

Effect sizes between	d_{corr} posttest	d_{corr} retention test
Conditions A -B	0.86	0.46
Conditions A -C	0.25	0.96
Conditions B -C	0.39	0.39

Table 6: corrected effect sizes between the three conditions

Students' experiences

Means and standard deviations on instructiveness, difficulty, and appreciation of the student questionnaire results showed that differences between the three conditions are minimal. In general all students thought they had learned from Word Score 2 and appreciated Word Score 2 positively, especially the students in conditions A and B. None of the students thought Word Score 2 was either too difficult or too easy, although results show a light tendency toward easy, especially for students in condition C.

Teachers' experiences

Overall, teachers that participated in the interviews (N=8) were positive about Word Score 2. All teaches thought Word Score 2 fit their vision on education and all were able to fit Word Score 2 in their practice. Most teachers (N=7) think that Word Score 2 can replace the conventional vocabulary method (except school B1). All teachers agree that playing Word Score 2 contributed to an expanded vocabulary of their students. Three teachers thought that the motivation of the students remained constant (school A1, C1, C2), two thought the motivation increased (school A3, B2) and two thought the motivation for playing the game decreased (school A2, B1). The latter was appointed to technical problems (A2) and lack of variation in the mini games (B1).

All teachers in conditions A and C where enthusiastic about the corresponding instruction: the usability was rated high and the difficulty was just right. The difficulty of the game was just right according to seven teachers; one thought the games (paper version) were too easy (school C1). The usability of the game was rated average by most teachers, except for two teachers (school B1, C2) who thought the game's usability was high. There were large differences in the time teachers put in preparation, varying from zero minutes (school B1) to 90 minutes (school A1), however all teachers indicated that this was as expected or that the preparation time was less than they had expected. In practice, most teachers spend 30 minutes a week on instruction gave their students 30 minutes a week for playing the game

Conclusions and Discussion

The effect of a serious game on students' vocabulary was investigated in a study containing three conditions: (A) Schools where students received vocabulary instruction and played the Word Score 2 game, (B) Schools where students did not receive the vocabulary instruction and where students played the Word Score 2 game and (C) Schools where students received vocabulary instruction and where students trained their vocabulary by playing a paper version of the Word Score 2 game.

Results show significant learning effects on the posttest for conditions that used the instruction next to the game. This could indicate that the instruction is an important part of Word Score 2 and needed when learning with serious games, which is corresponding to notions of Verhallen & Verhallen (1994) about pedagogies of vocabulary education. However, results show significant learning effects for all three conditions when comparing the retention test to the pretest. An explanation for this unexpected effect could be that students in condition B continued playing Word Score 2, although in a different theme, which may have indirectly led recalling words from the Halloween theme. Additionally students from condition B started with a significantly lower vocabulary level then the other two conditions. Although corrected for this different starting situation through analysis of covariance, it cannot be ruled out that specific instructional settings lead to higher or lower learning effects for students with a certain initial level. Further research concerning these aspects is needed. Finally, no differences were found between sexes.

Both teachers and students were very enthusiastic about Word Score 2. Students in all three conditions agreed Word Score 2 is instructive and fun, especially the students in the conditions that played the digital version of the game. According to the teachers, playing Word Score 2 contributed to an expanded vocabulary of their students. Furthermore, all teachers thought Word Score 2 fit their usual program and would be willing to replace the conventional vocabulary method.

References

Boyan, A. & Sherry, J.L. (2011). The challenge in creating games for education: aligning mental models with game models. *Child development perspectives, 5*(2), 82-87.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37- 46.

Cöp, J. (2009). *Hoe ICT het woordenschatonderwijs kan verbeteren*. Utrecht: PO Raad.

Hirumi, A., Appelman, B., Rieber, L. & Van Eck, R. (2010). Preparing instructional designers for game-based learning: part 2. *TechTrends, 54*(4), 19-27.

Kebritchi, M. & Hirumi, A. (2008). Examining the pedagogical foundations of modern educational computer games. *Computers & Education, 51*, 1729-1743.

Leemkuil, H.H. (2006). *Is it all in the game?: learner support in an educational knowledge management simulation game*. Enschede: University of Twente.

McCardle, P., & Chhabra, V. (2004). *The Voice of Evidence in Reading Research*. Baltimore: Brookes.

Peterson, M. (2010): Massively multiplayer online role-playing games as arenas for second language learning. *Computer Assisted Language Learning, 23*(5), 429-439.

Van Rooij, A. J., Jansz, J., & Schoenmakers, T. M. (2010). *Wat weten we over … effecten van games. Een beknopt overzicht van wetenschappelijk onderzoek naar de effecten van games* [What do we know about…. effects of games. An overview of scientiefic research into the effects of games]. Zoetermeer: Stichting Kennisnet.

Verhallen, M., & Verhallen, S. (1994). *Woorden leren, woorden onderwijzen* [Learning words, teaching words]. Hoevelaken: CPS.