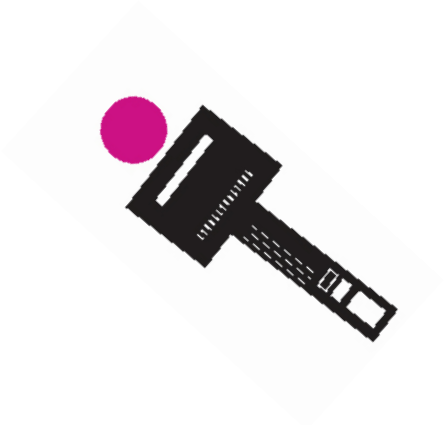
UNIVERSITY OF TWENTE.



TESTING AND FEEDBACK IN DGL

Faculty of Behavioural Sciences

Educational Science and Technology

Daian Geivid Saman Yan
s1189735
MSc. in Educational Science and Technology

November, 2013

Supervised by
Dr. H. H. Leemkuil
Dr. H. van der Meij

The Effect of Testing and Feedback in Digital Game-based Learning

Daian Geivid Saman Yan
S1189735

Supervised by Dr. H.H. Leemkuil & Dr. H. van der Meij

University of Twente
Faculty of Behavioural Sciences
Educational Science and Technology

Table of Contents

List of Tables

Abstract

Research on digital game-based learning indicates that this learning method can be - under certain conditions - more efficient than other traditional teaching methods, such as textbooks, or web-based learning. Without additional support learning from games may or may not occur, this is why instructional support should be used in educational settings. The present study focuses on testing with feedback as a type of instructional support with great promise. Testing is widely used in almost all classrooms around the world and there are plenty of studies supporting the benefits of testing and feedback in learning; however, there is scarce research about the use of testing and feedback as support in learning with games. This study examines the potential of testing and feedback to enhance game-based learning. Three conditions of instructional support are evaluated in an experiment: no instructional support, testing without feedback, and testing with feedback. In the study 48 international students from University of Twente played a strategy game, EnerCities, after their knowledge was tested and feedback was provided (depending on the condition). Later, there was another round of play and conclusively a final knowledge test. The findings show that participants in the testing and feedback condition improved their game and level score in the second round of game play.

Keywords: Digital game-based learning, learning from games, testing, feedback.

1 Introduction

The purpose of this study is to ascertain the effect of instructional support in digital game-based learning. The analysis is focused on testing and testing with feedback as types of instructional support. In the second chapter, an overview about the benefits and challenges of digital game-based learning is given; additionally, the rationale for choosing specific types of instructional support is explained and the conceptual framework is rationalized. The potential advantages of using digital games in classrooms are extensive and varied. However, there are limitations that need to be addressed in order to get out the most of them. Given the native nature of learning in games and the increasing influence of digital games in our society, integrating digital games in schools seems a logical idea; nonetheless, such integration is not a magical solution to enhance learning automatically.

The use of digital games in the classrooms is meant to improve learning outcomes and to achieve instructional objectives; in order to accomplish such goals instructional support is needed. In the third and fourth chapter the method and design of the study are described in detail, including the explication of the participants, materials, procedure, and analyses. Finally, in the last two chapters the results of the study are elucidated and the discussion about the limitations, scope, and lines of future research are set.

2 Definition of Key Terms

Digital Game-based Learning: Gee and Prensky (as cited in (Felicia, 2011a) made the term Digital Game-Based Learning (DGL) popular. DGL has been used to teach, train, and raise person's awareness and they may include one or all of these characteristics: game format, educational objectives, multimodal representations, feedback mechanisms, information provided to users, tools to track users' knowledge and proficiency, and adaptive pedagogical mechanisms (Felicia, 2011a).

Serious Games: A widely accepted definition of serious games is *"Any meaningful use of* ***computerized*** *game/game industry resources whose* ***chief*** *mission is not entertainment"* (Sawyer, 2007). Educational games are the predecessors of serious games (Djaouti, Alvarez, Jessel, & Rampnoux, 2011).

Due to the nature of this study these concepts are used interchangeably: digital game-based learning (DGL), serious games, educational games and digital games.

3 Testing and Feedback in Digital Game-based Learning

3.1 Problem Description

The use of digital game-based learning has three important advantages; first, games are a natural way of learning, that is, learning is genuinely pleasurable for human beings (Gee, 2007); second, digital games are increasingly popular, they are a growing part of our culture (Kirriemuir & McFarlane, 2004); and third, research has shown that DGL is more efficient than other teaching methods under certain conditions, such conditions include the use of instructional support as debriefing, collaboration, feedback and other types cited in Table 1 (Felicia, 2011b; Ke, 2009; Sitzmann, 2011; Wouters & van Oostendorp, 2013). The instructional support seeks to address problems in the use of DGL. One disadvantage is that educational games by themselves are not enough to improve learning, an inconvenience of DGL is its complexity, learners may get confused and overlook relevant information, and the main problem while using serious games is that students may develop misconceptions (Leemkuil & de Jong, 2012). In order to eliminate or minimize such problems instructional support is needed (Hays, 2005; Leemkuil & de Jong, 2012; Wouters & van Oostendorp, 2013).

Playing is a typical form of learning, it is the most old and time-honored instrument for education: game-playing has a vital educational function in mammals and birds, they learn by doing in a safe way, it is better to make mistakes while playing to hunt with siblings than with a true prey (Crawford, 1984). In nature, games are a learning tool, a tool that makes learning enjoyable. Nowadays, thanks to technology, digital games are an influential part of most children's leisure lives and progressively an significant part of our culture (Kirriemuir & McFarlane, 2004), people are spending more time playing games, only in US the number of gamers increased 241% from 2008 to 2011 (Macchiarella, 2012); not to mention that the use of digital games in learning and instruction has increased (Wouters, van der Spek, & van Oostendorp, 2011), due to the fast growth of usage of digital games, attention has been renewed to the role of games in education and to scientific research with focus in the design of educational games (Leemkuil, 2006). Recent studies have proved that digital games can be more effective than traditional instruction (Ke, 2009; Sitzmann, 2011; Wouters & van Oostendorp, 2013), researchers have found that digital games can be used, under certain conditions, to increase learning outcomes better than other media or teaching methods, such as text books or web-based learning (Felicia, 2011b).

Digital game-based learning can be more efficient than other traditional teaching methods (textbooks, web-based learning) under certain conditions, such as the use different types of instructional support.

3.2 Conceptual Framework

3.2.1 The process and vicissitudes of learning with games

According to Leemkuil and de Jong (2012), inquiry learning in simulation games is constituted by characteristic learning processes: orientation or exploration, generation of different possible solutions and hypothesis, appraisal of those options (experimentation), evaluation (reflection on the learning process) and monitoring of the outcomes. Unfortunately, de Jong (2006) sustains that learners may have difficulties with each of these learning process.

3.2.1.1 Orientation or Exploration

When playing digital games, pupils go through the orientation or exploration process, where they identify variables and their relationships (de Jong, 2006); in this process learners may encounter problems, because they may overlook critical data (Leemkuil & de Jong, 2012), or do not know what is relevant and consequently pay attention to irrelevant information by choosing wrong variables (Wouters & van Oostendorp, 2013), or in the worst case they may be overwhelmed by all the information that needs to be processed (Wouters et al., 2011); that is, they may experience cognitive overload (Leemkuil & de Jong, 2012).

3.2.1.2 Generation

The generation process is where pupils formulate a statement or a set of statements (de Jong, 2006), these statements can be seen as hypothesis or strategies; during this stage students may find hard to state testable hypothesis, they may try to achieve a state in the game rather than test strategies (Leemkuil & de Jong, 2012).

3.2.1.3 Limited or No Experimentation

In the experimentation process learners change variable values, make predictions, and interpret outcomes (de Jong, 2006); in this process students may design ineffective experiments by varying too many variables at one time, therefore they fail to make predictions and to interpret data correctly (de Jong, 2006); it is in this phase when novices implement learning process incorrectly (Leemkuil & de Jong, 2012) or develop misconceptions at the expense of activities that generate learning (Wouters & van Oostendorp, 2013).

Games, unlike simulations, do not provide much freedom for experimentation, this is due mainly to the restrictions of limited resources and the difficulty of recovery from wrong choices (Leemkuil, 2006; Leemkuil & de Jong, 2011); consequently while playing learners may or may not go through the experimentation process. However, trial and error may be used in the experimentation process to

overcome the limitations and difficulty of recovery, students may re-start the game several times with different settings by changing the variable values, they can do so until they succeed and win the game or stop trying. This trial and error technique can be compared with the experiential mode suggested by Leemkuil (2006), because eventually it may lead to learning, pupils may be able to know how to apply the knowledge after several trials, however they may not be able to explain it.

3.2.1.4 Appraisal or Evaluation

The appraisal process is where learners evaluate the validity of the hypothesis or the effectiveness of the possible solutions (de Jong, 2006); if since the beginning the hypothesis, strategies or possible solutions were designed incorrectly, in other words, they were not made to being assessed, then students may encounter problems like failing to make predictions or making mistakes when interpreting data (Leemkuil & de Jong, 2012).

3.2.1.5 Monitoring or Planning

The monitoring process is where learners examine the outcomes of what they have done or maintain an overview of what they have learned (de Jong, 2006; Leemkuil & de Jong, 2012), while the planning process is when students outline a schedule of the learning process (de Jong, 2006). Unfortunately, pupils tend to do only short-range planning and do not adequately monitor what they have done (Leemkuil & de Jong, 2012).

3.2.1.6 Reflection

The reflection process is when pupils reflect on their own behaviour and on the reaction of the system or other players to this (Leemkuil & de Jong, 2012). This process of reflection is defined by Leemkuil (2006) as the *reflective* mode that students use while learning in a digital game; this mode is where information processing is based on the use of learning strategies that can be supported by the aid of additional tools or other people (as other players and teachers).

3.2.2 The two types of learning in games

The previous learning processes can be identified in a more explicit model of how learning occurs in (serious) games. Koops and Hoevenaar (2012) propose the Serious Gaming Lemniscate Model (SGLM), depicted in Figure 1, to explain the two states involved in learning by serious games, the model suggests two cycles, a game cycle and a learning cycle. In the game cycle a student advances and discovers the rules of the game intuitively, (s)he acquires spontaneous (game induced) conceptual knowledge, only applicable within the game context - this cycle can be compared with what Leemkuil (2006) calls the experiential unselective mode. Whereas in the learning cycle the student systematically reflects, analyses, and examines the in-game experience, (s)he gains a more

formal or scientific conceptual understanding of what happens in the game that can be applied within the context of the game or in a completely different context - this cycle is what Leemkuil (2006) describes as selective reflective mode.

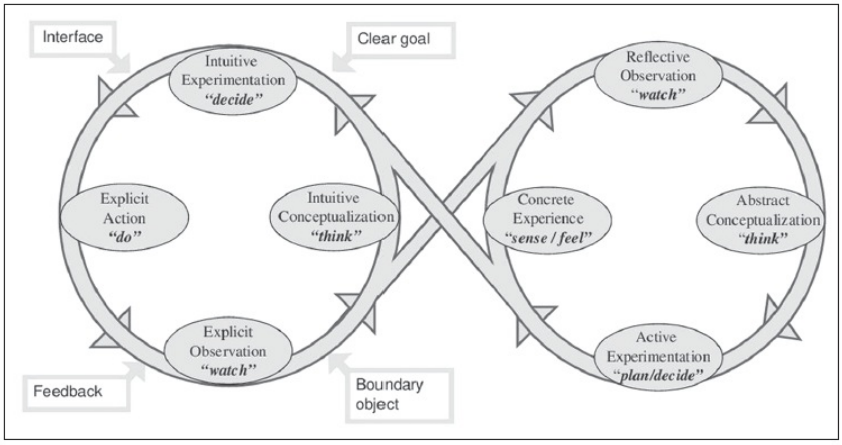


Figure 1: The Serious Gaming Lemniscate Model

According to SGLM a challenge with a serious game experience is to attract students out of the game cycle in order to rationally reflect on a scientific level (Koops & Hoevenaar, 2012). One of the tools to facilitate a *reflective* mode or learning cycle is instructional support. Recent studies agree that games are complex learning environments in which learners need instructional support to engage in cognitive processes (Leemkuil, 2006; Leemkuil & de Jong, 2012; Wouters & van Oostendorp, 2013); in other words in order to improve the learning outcomes with digital games, it should be supplemented with other instructional methods (Hays, 2005) to attract students into a learning cycle where they can reflect on a scientific level.

3.3 Rationale

Leemkuil (2006) proposes a model of game-based learning, which distinguishes two learning modes: experiential and reflective; the *experiential* mode is where learning is data-driven and reactive, students learn inadvertently specific things, in the form of facts, procedures and examples; when new abstractions or insights are learned in this way it is as implicit or intuitive knowledge, which is difficult to verbalize and transfer to other contexts. The *reflective* mode requires more effort than the *experiential*, in this mode information processing is based on the use of learning strategies that can be supported by the aid of additional tools or other people (as other players and teachers).

Additionally, Koops and Hoevenaar (2012) describe two types of cycles occurred within the SGML model: the game cycle and the learning cycle. During the game cycle reflection-in-action occurs in an intuitive sense; it develops a spontaneous concept that is only applicable within the context of the game, while during the learning cycle reflection-on-action consists of the ability of consciously and systematically reflect on the completed task, producing knowledge that can be used within the

context of the game or in other new different contexts outside the game. Reflection-in-action occurs while playing the game, however reflection-on-action is something that need to be induced, the learner should be pulled from the game cycle into the learning cycle (Koops & Hoevenaar, 2012).

These two approaches are very similar, although they may differ in name, what Leemkuil (2006) named as the *experiential unselective mode*, Koops and Hoevenaar (2012) called the *game cycle* as a form of intuitive, spontaneous learning that can be used only within the context of the game; and the concept of *selective reflective mode* (Leemkuil, 2006) is very similar to the *learning cycle* as a form of formal scientific learning that can be used within and outside the context of the game, in other words, learning can be transferred to other domains.

3.3.1 Instructional support in DGL

Instructional support can be used to induce learners into a reflective mode or into the learning cycle. There are extensive types of instructional support, and depending on their purpose they can be classified in different manners; a classification of types of instructional support in DGL based on recent studies is provided in Table 1. For example, Wouters and van Oostendorp (2013) enumerate different types of instructional support in their review; moreover, they classify 10 groups of diverse types of instructional support that are regularly used in digital games (reflection, modelling, advice, collaboration, interactivity, narrative elements, modality, feedback, personalization and others), Moreno and Mayer (2005) focus only in three types (guidance, reflection and interactivity), Johnson and Mayer (2010) concentrate only in one form of support (self-explanation), while Hays (2005) recommends two instructional methods (debriefing and feedback).

Among all the types of instructional support, there is one that can be found in several studies and reviews of DGL, this special category is often called reflection that is when learners are encouraged to think about their answers and explain it to themselves (Wouters & van Oostendorp, 2013) or to explain why an answer is correct (Moreno & Mayer, 2005; van der Meij, Leemkuil, & Li, 2013); it can be known as self-explanation too, when learners are induced to explain to themselves as they study (Johnson & Mayer, 2010). In any case, in order to be effective, reflection needs to be used in specific settings; Moreno and Mayer (2005) recommend to request students to reflect on correct information and models, rather than ask them to reflect on their own solutions and explanations; in like manner, Johnson and Mayer (2010) advise to induce the use of a selection format, where learners are given pre-formulated options to choose from, instead of a generation format, when learners are asked to compose their reasons. Although both articles study dissimilar questions and provide different arguments, they both recommend avoiding the practices of encourage pupils to reflect on explanations that may be wrong.

3.3.2 Testing and feedback to promote learning

Koops and Hoevenaar (2012) stimulate the switch from *game cycle* to *learning cycle* (a shift from *experiential* to *reflective* mode) by increasing the difficulty of the game to encourage students to look for help after they are confronted with a game-over screen. A different way to encourage and pull students from the game cycle or experiential mode to the learning cycle or reflective mode is through testing and feedback.

Testing can be seen as a switch from the game cycle to the learning cycle, by encouraging learners to think about their answers and explain it to themselves, to induce them into a reflective mode. In this case, testing and feedback is being used as a form of instructional support. Testing should be done not only with "why" interrogations, but also with questions that require students to think, deliberate or explain the implications of their answers; nevertheless, the explanations made by the pupils should be based on correct information. Therefore, learners should be provided with feedback, especially when their answers are not correct.

There are several studies that state the benefits of testing and feedback: Teachers can maximize the effective use of games by means as testing, feedback and replay (van der Meij & Leemkuil, in press). Include testing as part of a course might be an efficient strategy to improve learning outcomes (Kromann, Jensen, & Ringsted, 2009). Taking a test on material previously studied stimulates subsequent learning and retention of that material on a final test (McDaniel, Anderson, Derbish, & Morrisette, 2007). Additionally, Kromann et al. (2009) affirm that assessments that include feedback may induce learning and McDaniel et al. (2007) found that learning and retention were better when students were given feedback.

A mediated effect of testing is to help students to guide their future study toward material they have not yet mastered (Roediger & Karpicke, 2006); in other words, testing may help pupils to focus their attention on relevant information they need to learn. Studies suggest that testing helps to focus attention during game replay, making testing a good candidate for enhancing learning from games; the initial round of play has a more orientating function thanks to this players are more likely to focus on fundamental game features in the second round of play (van der Meij & Leemkuil, in press).

Testing can be used in two forms: as selection of relevant data, focusing the pupils' attention on significant information and in virtue of reflection, posing adequate questions to students. This duality of testing can be mapped on the cognitive architecture, testing can be used to help pupils in two ways; first, helping to search, select and focus on critical material making use of the working memory, and second, organizing, integrating and reflecting to structure new information into schemas in the long term memory.

In summary, testing, a widely known type of instructional support can be used for selection of relevant data and as switch from the game cycle or experiential mode into the learning cycle or reflective mode, whether feedback may be used to guide the reflection of students on correct information. In this study testing will be used as a switch into the reflective mode or learning cycle and feedback to support reflection on correct information, based on the recommendations of Johnson and Mayer (2010) and Moreno and Mayer (2005) to reflect on correct information instead of reflecting on self-generated information.

Author(s)	Types of Instructional Support
Hays (2005)	▪ Debriefing (reflection as steps to debrief) ▪ Feedback
Johnson and Mayer (2010)	▪ Self-explanation
Moreno and Mayer (2005)	▪ Guidance ▪ Reflection ▪ Interactivity
Wouters and Van Oostendorp (2013)	▪ ***Reflection:*** Reflection, self-explanation, elaboration, assignments. ▪ ***Modelling:*** Different types of scaffolding, modelling, worked examples. ▪ ***Advice:*** All types of advice whether contextualized, adaptive or not. ▪ ***Collaboration:*** Players played in dyads, groups or engaged in-group discussion. ▪ ***Interactivity:*** Interactivity, learner control and choice of game features. ▪ ***Narrative elements:*** Fantasy, rich narrative, foreshadowing, surprising events. ▪ ***Modality:*** Modality ▪ ***Feedback:*** Feedback, guidance. ▪ ***Personalization:*** Personalization, personalized messages. ▪ ***Others:*** Goal direction, pre-training, background information, cues, adaptivity.

Table 1: Classification of types of instructional support in digital game-based learning research.

3.4 Scientific relevance

There are several studies with focus on different types of instructional support to enhance learning from games (see Table 1), however, in none of those studies testing is addressed. There is little research on using testing on DGL, furthermore in general there is hardly any research concerning any type of instructional support in DGL (Leemkuil & de Jong, 2011). Educational research has virtually ignored testing as a technique to improve learning (McDaniel et al., 2007); dynamic testing (testing with feedback) does not only have a direct effect on learning, but also may encourage students to study more, to experience less test anxiety, and probably to even score better in standardized tests (Roediger & Karpicke, 2006). At the same time testing is widely use in educational settings as a form of summative evaluation, teachers and students are familiar with them, we suggest another form of evaluation, formative assessment that help to potentialize the benefits of testing.

4 Design of the Study and Research Questions

4.1 Design of the study

The purpose of this study is to ascertain whether testing and feedback affect learning from a game. Three conditions are compared: (1) no instructional support: no testing and no feedback, (2) testing without feedback, and (3) testing with feedback. Testing is done twice during the study, in both cases participants are asked to write down their scores and level at the game. First, all participants play the game and immediately after participants in condition two and three are tested for knowledge (test 1), the participants in the first condition receive a distractor, in other words they are not tested. Then participants in the third condition are provided with feedback on all answers, while participants in the second condition do no obtain any feedback. After, all the participants are invited to play another round of the game; the experiment is close with the second knowledge test (test 2). The overall experiment is depicted in Figure 2.

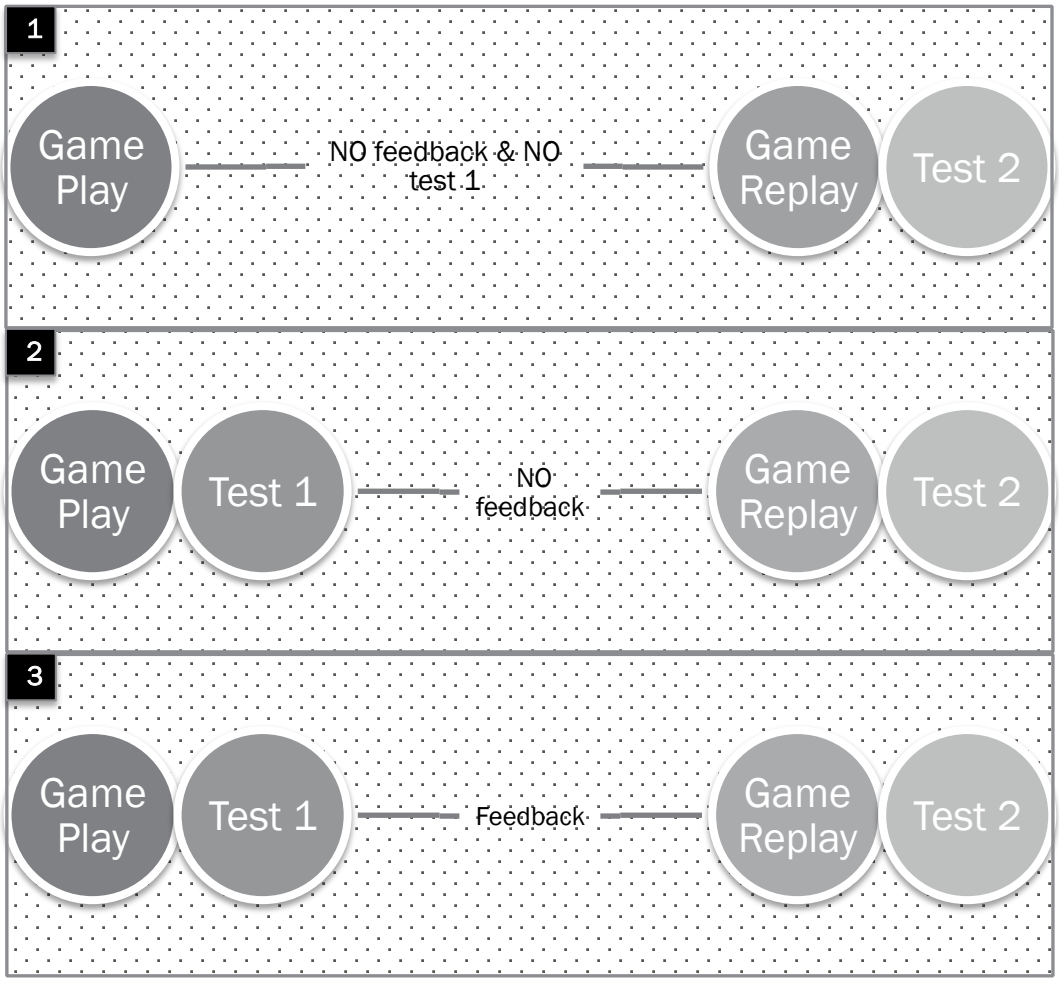


Figure 2: Conditions studied

4.2 Research Questions

The hypothesis is that learners with instructional support of testing and feedback would have the better learning outcomes (in games scores and test results) compared with the other two groups.

4.2.1 Question 1: Is there an effect of instructional support on game scores?

The prediction is that game scores in the second round of game replay will be higher than the first round of game play; it is likely that instructional support will be the main cause of the increase, in the case of the first condition where no instructional support is offered a limited increase is expected by effect of playing the game itself.

4.2.1.1 Sub-question 1A: Is there an effect of testing on game scores?

Testing is expected to improve moderately the game scores; it is probable that testing will make players to organize new information into cognitive structures and integrate such structures by building connections between prior knowledge (long term memory) and short-term memory.

4.2.1.2 Sub-question 1B: Is there an effect of testing and feedback on game scores?

Testing and feedback is expected to be the cause of significant increase on game scores; it is likely that testing and feedback will help students not only to organize and integrate new cognitive structures, but also to corroborate whether those structures or new knowledge is correct; as result it is likely that testing and feedback will have a higher effect on game scores compared to only testing.

4.2.2 Question 2: Is there an effect of instructional support on test scores?

The prediction is that participants who receive both types of instructional support, testing and feedback, will learn more than those who received only testing as instructional support; additionally, participants who did not receive any instructional support are expected to have an increase of limited learning after the round of game replay, but less compared to those with instructional support. The reasoning is that feedback after testing helps students not only to organize and integrate new information, but also to confirm that such new information is based on correct information, in this way students avoid the development of misconceptions; also, the round of game replay may increase the learning outcomes of students as suggested by Leemkuil and de Jong (2012) by allowing students to test the hypothesis they formed in the first round playing the game.

4.2.3 Question 3: Is there a relationship between game score and test results?

The relation between game scores and test results is analysed in different points in time. A positive correlation is expected at the end of the study, in other words in the game scores of game replay and test results on the final test. The rationale is that participants will be able to focus on relevant game features, or test hypothesis in the game replay, whether the first found of game play is probable to have a familiarising function to form such hypothesis.

4.2.4 Why evaluate games scores?

In DGL is important to evaluate the game scores, they are by default an indicator of how good students are doing in the game. Furthermore, game scores have construct validity, players must consider a variety of factors within the game in order to achieve a high score. Game scores are based on real time choices players need to face in the game; they reflect the results of decision-making. Another advantage is that game scores are an unobtrusive metric of the pupil's competences (van der Meij et al., 2013); for this reason a positive correlation between game score and test results is expected.

5 Method

5.1 Participants

Participants of the study were 48 international students from University of Twente. The age of students ranged between 19 and 33. Participants volunteered for the study, they had different backgrounds, (bachelor, master and PhD studies) and specialization areas, such as chemistry, mathematics, communication, history, psychology and others. They were randomly assigned to each of the three groups.

The native language of the participants was varied: Spanish, German, Hindu, French, among others; they have learned English for several years, their studies were imparted in English and they need to probe a minimum level of English: TOEFL iBT 80 points or IELTS 6.0 overall band score to be accepted at University of Twente. All students were requested to play the game in English.

5.2 Materials

The materials used in this study are game, game experience questionnaire, knowledge test one, feedback and knowledge test two. The questionnaire on game experience is used to test the difference between participants. All the materials were in English.

Outcomes were measured with game scores and paper-based knowledge tests; they were measured twice, one is after the first round of game play and the other one after the round of game replay; participants in the control condition receive a distractor instead of the knowledge test one.

5.2.1 Game

EnerCities is a serious game which main goal is to build the most sustainable city until it reaches 200 inhabitants. The students needs to manage the energy balance, cash reserve and natural resources in order to balance the economical (profit), environmental (planet) and wellbeing (people) scores while supplying the growing city with sufficient electricity, implementing energy conservation and CO2 emission measures and minimizing fossil fuel consumption to keep the planet healthy, at the same time the players need to increase the population, generate money and offer public services to keep the citizens happy and grow the city.

The game starts with a small community and a small piece of land to build on. A drag-and-drop interface allows players to build structures (e.g. residential and industrial areas, renewable / non-renewable energy sources, green zones) to expand the city (Knol, 2011). Each student's choice influences the scores for economy, environment and well-being. Players can reach until the forth level,

in each new level they receive more city space to expand it and extra available game options; each new level can be reached when the city gains a certain amount of population. The game is over when the population reach 200 citizens and you have a positive score.

	Indicators	Variables	Scores
Purpose	Aid to check the current state of:	To build structures of type:	To evaluate the influence of taken decisions on:
Options	▪ Energy balance ▪ Cash reserve ▪ Natural resources	▪ Residential: Increase population ▪ Economic: Produce money ▪ Environmental: Keep planet healthy ▪ Well-being: Keep citizens happy ▪ Energy: Power the city	▪ Economy score ▪ Environmental score ▪ Well-being score

Table 2: Indicators, variables and scores in the game EnerCities.

Due to the indicators, variables and scores the players need to pay attention to (summarized in Table 2), it can be a real challenging game. The main condition is the natural resources; this is the only indicator that cannot be increased by any strategy. The game start with 1000 points in natural resources and there is no way to generate new of them. The task is balance your economy, energy and well-being keeping the planet healthy and without running dry of oil. The evaluation of the players' performance is an overall high score automatically given by the game that reflects the economy score, the environmental score and the well-being score.

The game allows players to execute several strategies and see the results of their actions on the long term. The duration of the game is approximately 15-45 minutes, depending on the player's strategies (Knol, 2011).

5.2.2 Game Experience Questionnaire

The questionnaire on game experience consists of seven questions. Three questions are about the participants' previous experience in game playing, such questions inquire the time spent on playing digital games, playing strategy games and playing EnerCities (e.g. In average how many hours per week do you spend on playing games?) Answers are given in predetermined categories for ranges of hours, from "None" to "More than 10 hours". The first two questions are open questions related to the possibility of previous knowledge about environmental or energy saving topics associated to students' studies or hobbies (e.g. Do you have studies or hobbies related with environmental or energy saving? If yes, which ones?) The two final question are about the students' level of English, one question is answered in a 6-point scale based on the Common European Framework of Reference for Languages ranging from "Basic User: Beginner" until "Proficient User: Mastery". The Appendix A shows the game experience questionnaire.

5.2.3 Knowledge Test

The knowledge test is based on the content analysis of EnerCities, the test measures knowledge about facts, concepts, principles, and structures in the game.

The knowledge test consists of 10 questions, 2 questions about facts, 2 questions about concepts, 2 questions about principles and 4 questions about structures. The answers to each question are scored as correct (between 1 and 3 points) or incorrect (0 point); the answers are evaluated based on units of meaning that should be present in the answer. The maximum possible score for the knowledge test is 19 points. Table 3 summaries the composition of the test scoring.

Questions about facts interrogate about specific information that does not change during the game (e.g. "What are the advantages and disadvantages of a nuclear plant?"). The maximum score for all facts questions is 5 points.

Questions about concepts inquire about descriptions, depictions, or definitions of events on the game such as power consumption (e.g. "How can you improve the score of this icon ⚡ in the game?"), this is not explained in the game but represents implicitly the energy balance between generation of electricity and power consumption. Players receive points for concepts that are not explicitly defined in the game, rather students must infer their meaning from playing the game. The maximum score for concept questions is 4 points.

	Type	Maximum points
1	Facts	3
2	Structure	2
3	Facts	2
4	Structure	3
5	Concepts	3
6	Structure	2
7	Concepts	1
8	Structure	1
9	Principle	1
10	Principle	1
	TOTAL	19

Table 3: Test scores composition

Questions about principles ask about the influence of the participants' actions, events and outcomes between each other (e.g. "How can a super solar plant provide a higher amount of energy than a traditional solar plant?"). The maximum score for principle questions is 2 points.

Questions about structures refer to the coherence and relationships between the various principles of the game. Such questions consists of two parts, first they described a situation or an event through the use of figures with specific conditions on the game (e.g. the environmental score is negative), the second part ask players which specific actions they should take and/or the possible effects of such actions (e.g. "What steps would you take to solve this problem?"). The maximum score for structure questions is 5 points.

There are two similar tests (version A and B); the items of both knowledge tests are comparable, different in content but parallel in type, number of items and process of scoring. See Appendix B and C for the knowledge tests.

5.2.4 Rubric

Rubrics were used in order to delineate consistent criteria for grading the knowledge tests; two rubrics were created, one for each test version. The rubrics were developed in a form of a table where each row was a question and the columns contained the maximum points awarded and the description of the best answer and acceptable variations of the answers. An example of the rubrics used to asses the knowledge tests can be seen in the Table 4. See Appendix D and E for the rubrics.

Q	Evaluation	3 point	2 point	1 point
#	Possible explanations of why choosing the **B** option (improved insulation): - The natural resources ( ) cannot be regenerated - This option would represent the highest saving of natural resources.	Not applicable	The option B was selected. The answer includes at least 1 explanation from the evaluation section.	The option B was selected.

Table 4: Example of the rubrics used to score the knowledge tests.

5.2.5 Feedback

The feedback given to students consist of all the answers for facts, concepts and principles based on the information in the EnerCities game; only in the case of structure knowledge questions students receive the best possible options to take under the depicted conditions. In the Table 5 an example of feedback provided to students is shown. The feedback was given after the test in the form of knowledge of correct results (KCR) plus additional information. See Appendix B and C for the feedback provided to students.

Which three energy sources supply the most energy?
1. Nuclear fusion plant (100 points) 2. Super Solar (50 points) 3. Nuclear Plant (40 points)

Table 5: Example of feedback for facts questions.

5.3 Procedure

The experiment was conducted with clusters of students and their laptops grouped in tables. A brief description of the procedure emphasizing that no help could be provided to them in case of questions about the game was given; following the subjects completed game experience questionnaire. Then participants could play the game individually by 20 minutes maximum; afterwards subjects of the groups of instructional support were individually tested for knowledge (test 1), participants in the control condition were allowed for a break of 20 minutes so they could surf on the web. When the subjects completed the test they were requested to handle back the test and participants in the

testing and feedback condition were required to read the printed feedback that included the questions and correct answers to each of them, while participants in the testing group were allowed for a break of 5 minutes to surf on the web. Then participants played the game individually again for another 20 minutes; the experiment was finalized with another individual test (test 2).

5.4 Analyses

Kruskal-Wallis tests were conducted to evaluate whether there was significant difference between the three groups in game experience (general digital games, strategy games and EnerCities), a Kruskal-Wallis test was run to assess the random distribution of participant across conditions regarding the English level and a Chi-square (χ^2) test for association was used to calculate whether there was significant difference between the three groups in interest in environment and energy saving. The analyses showed that there was not statistically significant difference between conditions for any type of game experience; likewise, there were no statistically significant differences in the English level of participants.

Paired-sample t-tests were used to assess whether there was significant difference between the game score of the first round compare to the second round for the three conditions. Also, Wilcoxon Signed-Rank tests were used to evaluate the presence of significant difference between the game levels of the rounds of play for the three groups. A mixed ANOVA test was used to evaluate the effect of time and interaction between time and condition.

Two outliers were detected in the group of testing and feedback, they were more than 1.5 box-lengths from the edge of the box in a boxplot, inspection of its values did not reveal data entry or measurement errors, to avoid misinterpretation the outliers were removed.

A mixed ANOVA analysis was used to evaluate the effect of two different instructional support conditions on the scores.

A Pearson's product-moment correlation was run to assess the inter-rater reliability for the tests. Preliminary analyses showed the relationship to be linear with both variables normally distributed, as assessed by Shapiro-Wilk test ($p > .05$), and there were no outliers. There was a strong positive correlation between the raters, $r = .889$, $p = .044$ for the test version A, and $r = .910$, $p = .032$ for the version B. In other words, the grading of the tests was similar despite the grading of two different raters.

6 Results

6.1 Do the groups differ in background?

Kruskal-Wallis tests were conducted to evaluate whether there was a significant difference between the three groups in game experience (general digital games, strategy games and EnerCities), and to assess the random distribution of participant across conditions regarding the English level. A chi-square (χ^2) test for association was used to calculate whether there was significant difference between the three groups in interest in environment and energy saving. The Table 6 summarizes the mean and frequencies in game experience, interested in environment and English level of participants.

	No Instructional Support				Testing				Testing and Feedback			
	None	0-5	6-10	>10	None	0-5	6-10	>10	None	0-5	6-10	>10
Game Playing Time	6	8	2	0	5	10	1	0	11	2	1	0
	Mean rank = 26.13				Mean rank = 26.69				Mean rank = 20.69			
Strategy Game Playing Time	14	2	0	0	12	4	0	0	14	0	0	0
	Mean rank = 23.44				Mean rank = 26.38				Mean rank = 23.69			
EnerCities Playing Time	16	0	0	0	16	0	0	0	14	0	0	0
	Mean rank = 24.50				Mean rank = 24.50				Mean rank = 24.50			

Interest in environment (energy saving)	Yes	No	Yes	No	Yes	No
	2	14	3	13	8	6

English Level	Beginner	Elementary	Intermediate	Upper Intermediate	Operational Proficiency	Mastery	**Beginner**	**Elementary**	**Intermediate**	**Upper Intermediate**	**Operational Proficiency**	**Mastery**	Beginner	Elementary	Intermediate	Upper Intermediate	Operational Proficiency	Mastery
	0	0	**2**	2	8	4	0	0	0	4	7	5	1	1	1	3	7	1

Table 6: Summary of mean and frequencies in game experience, interest in environment and English level.

The first question is about participants' average hours per week spending on playing games. It was designed to be answered with four ordinal values, ranging from none, between 0 to 5 hours, between 6 to 10 hours, and more than 10 hours. The higher score represents the more hours participants spent on playing games. The time of playing games score was not statistically significantly different between the different types of instructional support, $\chi^2(2) = 3.080$, $p = .214$; in other words there is no significant difference in the hours spent on playing games between the three groups.

Participants' average hours per week spending on playing strategy games is related to the second question. It was designed to be answered with four ordinal values, ranging from none, between 0 to 5 hours, between 6 to 10 hours, and more than 10 hours. The higher score represents the more hours participants spent on playing strategy games. Again, the time of playing strategy games score was not

statistically significantly different between the different types of instructional support, $\chi^2(2) = 1.943$, $p = .379$; in other words there is no significant difference in the hours spent on playing strategy games between the three groups.

The time spent playing EnerCities was measured with the same four ordinal values. All participants from the three different groups never played the game before; consequently there is no difference between the groups.

The English level of participants was asked based on the Common European Framework of Reference for Languages with a nominal variable with six values: beginner, elementary, intermediate, upper intermediate, operational proficiency and mastery. The English level was not statistically significantly different between the different types of instructional support, $\chi^2(2) = 3.540$, $p = .170$; in other words there is no significant difference in the English level between the three groups.

The interest of participants related with environmental or energy savings could be answered with a dichotomous variable which values were yes or no. A Chi-square test for association shows that there is a statistically significant association between interest for the environment and the three groups, $\chi^2(2) = 8.434$, $p = .015$. There is a moderate association between interest for the environment (or energy saving) and the three groups, $\varphi = 0.428$, $p = .015$. However, a detailed analysis shows that there was no statistically significant correlation between the interest of participants in environmental or energy savings and the first and second round of game scores ($r = -.003$, $p = .982$; $r = 106$, $p = .483$) and the first and second knowledge tests ($r = .123$, $p = .519$; $r = -.009$, $p = .953$); therefore we can conclude that there is no significant difference in interest for the environment and the results of the game and test scores.

Summarizing, it is concluded that prior to game-play the three groups of participants are not different in game experience, English level and interest on environmental or energy savings.

6.2 Is there an effect of instructional support on game scores?

The differences between the game scores in the second and first round of play were normally distributed, as assessed by Shapiro-Wilk's test ($p = .075$ for no instructional support, $p = .956$ for testing and $p = .960$ for testing and feedback). A summary of means and standard deviations for game and level scores is shown in .

A mixed ANOVA displayed that there was an effect of time in game scores $F(1,43) = 4.292$, $p = .044$ partial $\eta^2 = .091$. Table 7 shows that game scores in the second round were higher than the first one.

	Game Score 1		Game Score 2		Level Score 1		Level Score 2	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
No Instructional Support	113.44	54.49	116.50	63.83	2.75	.931	3.13	.885
Testing	116.19	52.40	144.25	49.30	3.00	.816	3.31	.873
Testing and feedback	97.57	32.19	120.19	39.23	2.57	.852	3.21	.893
TOTAL	109.57	47.64	127.30	52.65	2.78	.867	3.22	.867

Table 7: Game and level scores for the first and second round.

6.2.1 Is there an effect of re-play without instructional support on game scores?

Paired-samples t-test was used to determine if there was significant change between the game scores in the first and second round of gameplay when there was no instructional support. Participants score better in the second round of gameplay (116.50 ± 63.83 points) as opposed to the first round of gameplay (113.44 ± 54.49 points). The second round of play showed an increase of 3.063 (95% CI, -23.47 to 29.59) points compared to the first round of gameplay. The second round of play did not display a statistically significant increase in points compared to the first round of play, $t(15) = .246$, $p = .809$.

6.2.1.1 Is there an effect of re-play without instructional support on level scores?

A Wilcoxon Signed-Rank test was run to determine if there were differences in level scores when subjects played in the first round of play vs. second round play when there was no instructional support. There was a statistically significant increase in level scores in the second round of play compared to the first round of play, $z = 2.121$, $p = .034$, $r = 0.37$.

6.2.2 Is there an effect of testing on game scores?

For the second group a paired-samples t-test was used to evaluate if there was significant change between the game scores in the first and second round of play after testing. Participants scored better in the second round of gameplay (144.25 ± 49.30 points) as opposed to the first round of gameplay (116.19 ± 52.40 points). The second round of play showed an increase of 28.06 (95% CI, -12.186 to 68.311) points compared to the first round of gameplay. The second round of play did not display a statistically significant increase in points compared to the first round of play, $t(15) = 1.486$, $p = .158$.

6.2.2.1 Is there an effect of testing on level scores?

For the second condition a Wilcoxon Signed-Rank test was run to determine if there was significant change between level scores in the first and second round of play after testing. There was not

statistically significant change in level scores in the second round of play compared to the first round of play, z = 1.184, p = .236.

6.2.3 Is there an effect of testing and feedback on game scores?

A paired-samples t-test was used to assess if there was significant change between the game scores in the first and second round of play after testing and feedback. Participants scored better in the second round of gameplay (120.29 ± 39.23 points) as opposed to the first round of gameplay (97.57 ± 32.19 points). The second round of play showed an increase of 22.71 (95% CI, -2.649 to 48.077) points compared to the first round of gameplay. The second round of play displayed a marginal significant increase in points compared to the first round of play, t(13) = 1.935, p = .075.

6.2.3.1 Is there an effect of testing and feedback on level scores?

A Wilcoxon Signed-Rank test was run to determine if there was significant difference between level scores in the first and second round of play after testing and feedback. There was a statistically significant increase in level scores in the second round of play compared to the first round of play, z = 2.310, p = .021, r = .44.

6.3 Is there an effect of instructional support on knowledge test?

A mixed ANOVA was run to determine the effect of conditions and time on the test scores. There were no outliers in the data, as assessed by inspection of a boxplot for values greater than 1.5 box-lengths from the edge of the box and by studentized residuals not greater than 3 standard deviations either. Test results were normally distributed for all interventions at all time points, as assessed by Shapiro-Wilk's test ($p > .05$). There was homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance ($p = .160$ for test score in round 1 and $p = .537$ for test score in round 2). There was homogeneity of covariances, as assessed by Box's test of equality of covariance matrices ($p = .840$). There was no statistically significant interaction between the conditions and time on test scores, $F(1,28) = .078$, $p = .782$, partial $\eta^2 = .003$. The main effect of time did not show a statistically significant difference in test scores at the different time points, $F(1, 28) = .197$, $p = .660$, partial $\eta^2 = .007$. That is, there was no improvement in test scores after the conditions. The shows a summary of the knowledge test scores means and standard deviations.

The knowledge test is composed of four types of questions: facts, concepts, structure and principles. Additional analyses with paired-samples t-test for each type of question between the knowledge tests were done in the experimental conditions. For each type of question the second knowledge test did not display statistically significant increase in points compared to the first knowledge test. Table 9 summarizes the means and standard deviations of test scores for each type of question.

	Test 1		Test 2	
	Mean	S.D.	Mean	S.D.
No instructional support	NA	NA	9.81	2.56
Testing	9.25	2.72	9.56	2.78
Testing and feedback	8.50	3.20	8.57	3.29
TOTAL	8.90	2.93	9.35	2.86

Table 8: Test scores for the first and second knowledge test.

		No instructional Support		Testing		Testing and Feedback	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
TEST 1	Facts	NA	NA	3.06	1.52	2.29	1.63
	Concepts	NA	NA	1.13	1.40	1.79	1.67
	Structure	NA	NA	4.13	1.62	3.64	1.21
	Principles	NA	NA	.75	.775	.79	.699
TEST 2	Facts	3.19	1.11	2.69	1.44	2.29	1.20
	Concepts	1.94	1.69	1.69	1.35	1.21	1.57
	Structure	3.94	1.53	4.31	2.05	4.21	1.05
	Principles	0.75	.683	.88	.619	.86	.864

Table 9: Test scores for each type of questions

6.4 Is there a relationship between game score and test results?

Analyzing the whole group of participants, and based on Cohen (1988), there was a moderate positive correlation between game scores in the first round and game scores in the second round, r = .390, p = .007 and a strong positive correlation between the two knowledge test scores r = .695, p < .0005. A small correlation was found between the second game score and the second test score with r = .296, p = .046 and a moderate correlation between the first round of game scores and the second test scores r = .485, p = .001.

Table 10 depicts the results from the correlational analysis for the group of no instructional support; two strong correlations were found one between the game scores and one between the first game scores and the test 2. Table 11 show the correlations for the groups with instructional support split for the two conditions; the testing condition showed two strong correlations, one between test scores and one between test 2 and the first game scores. In the testing and feedback group there was a strong correlation between the knowledge test scores.

	Game 1	Game 2	Test 2
Game 1	X	.759*	.607**
Game 2	X	X	.417
Test 2	X	X	X

Table 10: Correlations between test and game scores with no instructional support

*p = .001
**p = .013

	Game 1	Game 2	Test 1	Test 2
Game 1	X	-.106	.442	.554*
Game 2	*.451*	X	.204	.331
Test 1	*.195*	*.155*	X	.588**
Test 2	*.267*	*.168*	***.771****	X

Table 11: Correlation between test and game scores with instructional support

The underline cells above the X diagonal display the correlation for participants from the testing group (n = 16)
The italic cells below the X diagonal display the correlation for participants from the testing and feedback group (n = 14)
*p = .026
**p = .017
***p = .001

7 Discussion

The experimental findings reveal that testing and feedback have a positive effect in digital game-based learning to perform well in the game. The results revealed a significant improvement in game level scores, especially in the testing and feedback condition; all participants averagely increase 17.74 points in game performance, while the subjects in testing and feedback condition increase 22.71 points in game scores and showed a significant increase in the level scores. Although the condition of testing showed the highest increase with 28.06 points in game scores, the group did not show a significant increase in the level scores.

There was no effect of condition on the test scores. Participants with instructional support, testing and feedback and only testing, did not increase their test score compared to the group with no instructional support. Moreno and Mayer (2005) summarized their findings stating that reflection alone does not promote deeper learning unless it is based on correct information, in this study feedback was used to support reflection on correct information. The feedback was given after the test in the form of knowledge of correct results (KCR) plus additional information, this type of feedback provide knowledge of which is the correct answer and more information related to it; knowledge of correct results leads to better learning compared to other types of feedback as knowledge of results (van der Meij, Albers, & Leemkuil, 2011). However, the nature of EnerCities game may give an explanation for the absence of effect of testing and feedback on test scores. Moreno and Mayer (2005) suggested that reflection techniques help students to learn from non-interactive conditions but reflection in interactive environments may not significantly stir learning. EnerCities is a very interactive game and while students play the game they continuously face situations where they need to choose among various options. When players are required to make choices during the game their cognitive activity is already at a high level (Moreno & Mayer, 2005), with questions such as *"Why would you choose this option?"* players are not challenged because they already evaluated and analysed such choices during the game. Additionally, students may reflect on a self-generated wrong answer that may promote the consolidation of an incorrect mental model (Moreno & Mayer, 2005).

An interesting result is that the condition has a positive effect in digital game-based learning in game scores but not in test scores; (van der Meij & Leemkuil, in press; van der Meij et al., 2013) also found improvement in game scores over time, however contrary to the findings of this study, they did find improvement in test scores over time. An explanation could be that participants received feedback during the game with changes in game score immediately after they performed or choose certain actions, therefore they could evaluate and assess the effect of their movements while playing, while doing the tests participants did not receive immediate feedback, and the group that receive feedback had the opportunity to reflect on possible incorrect answers and later know that those were not the right options. Furthermore, students did not have the opportunity to compare their answers with the feedback, so they were not able to check which of their answers were correct to corroborate the

information and which answers were wrong. In other words, students with feedback may have reflected on wrong answers and afterwards may or may not know that their conceptions were erroneous. Another possible explanation is that immediate feedback in the game while playing may have promoted a self-competitive attitude in the participants and therefore helped them to increase their scores in the second round of the game.

The findings of this study corroborate the view of Moreno and Mayer (2005) that using elaborative interrogation or reflection before students get corrective feedback may promote the development of an incorrect mental model by making students articulate their misconceptions. These findings leave open the possibility to design another study where reflection through testing is done since the beginning: instead of asking to the subjects to choose the correct options in the test and explain why (where students may reflect on incorrect information if they choose a wrong answer); the correct answers should be given in the test and then ask to participants to explain why they think this is the correct choice; in this way students are required to reflect on correct information since the beginning and feedback would help to confirm or guide students when wrong misconceptions were developed. Furthermore, futures studies may consider give feedback in a different way, by allowing students to compared their answers in the test with the printed feedback. In this way the participants will be able to compare their wrong answers with the correct ones.

As expected, the correlations between the game scores and test scores in both rounds were significant for all participants in the condition, this finding indicate consistency: participants who scored well in the first round of play and testing perform similarly in the second round. The subjects in the No Instructional Support group showed a significant correlation between the game scores for the two rounds, while the participants in the Testing and Testing and Feedback condition showed a significant correlation between the test scores for the two rounds; it is unclear why the experimental conditions did not show a correlation between the game scores.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Vol. Second edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crawford, C. (1984). *The Art of Computer Game Design* (Vol. 2008): Osborne/McGraw-Hill.

de Jong, T. (2006). Computer Simulations - Technological Advances in Inquiry Learning. *Science, 312*, 532-533.

Djaouti, D., Alvarez, J., Jessel, J.-P., & Rampnoux, O. (2011). Origins of Serious Games. In M. Ma, A. Oikonomou & L. C. Jain (Eds.), *Serious Games and Edutainment Applications* (pp. 25-43): Springer London.

Felicia, P. (2011a). How can a digital game for learning be defined?: Waterford Institute of Technology.

Felicia, P. (2011b). What evidence is there that digital games can be better than traditional methods to motivate and teach students? : Waterford Institute of Technology.

Gee, J. P. (2007). *Good video games + good learning: Collected essays on video games, learning and literacy*. New York: Peter Lang.

Hays, R. (2005). The Effectiveness of Instructional Games: A Literature Review and Discussion. Orlando, FL: Naval Air Warfare Center, Training Systems Division.

Johnson, C. I., & Mayer, R. E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior, 26*(6), 1246-1252. doi: 10.1016/j.chb.2010.03.025

Ke, F. (2009). A Qualitative Meta-Analysis of Computer Games as Learning Tools. In R. E. Ferdig (Ed.), (pp. 1-32). Hershey, PA: IGI Global.

Kirriemuir, J., & McFarlane, A. (2004). Literature Review in Games and Learning. Bristol: FutureLab Series - Report 8.

Knol, E. (2011). Effects of serious game EnerCities on energy-related attitudes and behaviours.

Koops, M., & Hoevenaar, M. (2012). Conceptual Change During a Serious Game: Using a Lemniscate Model to Compare Strategies in a Physics Game. *Simulation & Gaming*. doi: 10.1177/1046878112459261

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical education, 43*(1), 21-27.

Leemkuil, H. (2006). *Is it all in the game?: Learner support in an educational knowledge management simulation game.*

Leemkuil, H., & de Jong, T. (2011). Instructional Support in Games. In J. D. Fletcher & S. Tobias (Eds.), *Computer Games and Instruction* (pp. 353-369). Charlotte, NC: Information Age Publishing.

Leemkuil, H., & de Jong, T. (2012). Adaptive Advice in Learning With a Computer-Based Knowledge Management Simulation Game. *Academy of Management Learning & Education, 11*(4), 653-665.

Macchiarella, P. (2012). Trends in digital gaming: Free-to-Play, Social, and Mobile Games. *Parks Associates,* 1-8.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4-5), 494-513.

Moreno, R., & Mayer, R. E. (2005). Role of Guidance, Reflection, and Interactivity in an Agent-Based Multimedia Game. *Journal of Educational Psychology, 97*(1), 117-128.

Roediger, H. L. I., & Karpicke, J. D. (2006). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science, 1*(3), 181-210.

Sawyer, B. (2007). *The "Serious Games" Landscape.* Paper presented at the Instructional & Research Technology Symposium for Arts, Humanities and Social Sciences, Camden, USA.

Sitzmann, T. (2011). A Meta-Analytic Examination of the Instructional Effectiveness of Computer-Based Simulation Games. *Personnel Psychology, 64*(2), 489-528. doi: 10.1111/j.1744-6570.2011.01190.x

van der Meij, H., Albers, E., & Leemkuil, H. (2011). Learning from games: Does collaboration help? *British Journal of Educational Technology, 42*(4), 655-664. doi: 10.1111/j.1467-8535.2010.01067.x

van der Meij, H., & Leemkuil, H. (in press). Do collaborative gaming and testing with replay enhance learning from games?

van der Meij, H., Leemkuil, H., & Li, J.-L. (2013). Does individual or collaborative self-debriefing better enhance learning from games? *Computers in Human Behavior, 29*(6), 2471-2479. doi: http://dx.doi.org/10.1016/j.chb.2013.06.001

Wouters, P., van der Spek, E. D., & van Oostendorp, H. (2011). Measuring learning in serious games: a case study with structural assessment. *Educational Technology Research and Development, 59*(6), 741-763.

Wouters, P., & van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Computers & Education, 60*(1), 412-425.

Appendix A: Game Experience Questionnaire

Name: _____ Age: _____

1. What are you studying?

2. Do you have studies or hobbies related with environmental or energy saving? (If yes, which ones?)

☐ Yes:_____ ☐ No

3. In average how many hours per week do you spend on playing games?

☐ None
☐ Between 0 – 5 hours
☐ Between 6 – 10 hours
☐ More than 10 hours

4. How many hours on average did you play a strategy game last week (like the Sims, SimCity, or Civilization)?

☐ None
☐ Between 0 – 5 hours
☐ Between 6 – 10 hours
☐ More than 10 hours

5. How many hours on average did you play "Enercities" game last week?

☐ None
☐ Between 0 – 5 hours
☐ Between 6 – 10 hours
☐ More than 10 hours

6. What level of English do you have?

☐ Basic User: Beginner
☐ Basic User: Elementary
☐ Independent User: Intermediate
☐ Independent User: Upper Intermediate
☐ Proficient User: Operational proficiency
☐ Proficient User: Mastery

7. Please write down the scores and type of test of the last English test you took e.g. "TOEFL iBT 76 points": _____

Appendix B: Test A with its feedback

Name: _____ Age: _____

Game Score: _____ At level: _____

Question 1

Which three energy sources supply the most energy?
1. Nuclear fusion plant (100 points) 2. Super Solar (50 points) 3. Nuclear Plant (40 points)

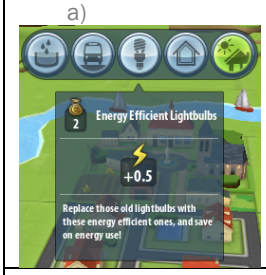
Question 2

When you click on a **Suburban building** the menu as depicted on the side will appear:

This player has already chosen to place a **solar roof**, by which he will gain 1 point for natural resources: , and he will gain 1 point for energy balance: .

He wants to improve this suburban building further and therefore has to choose between the four remaining options. Select the option that you consider best.

a)



b)

c)

d) THIS IS THE CORRECT ANSWER

Why would you choose this option?

The reserves of natural resources ( ) are the only resource that cannot be “regenerated” or increased by building any structure. Therefore natural resources may be used carefully and wisely: A suburban building uses 2 points of natural resources, with the solar roofs 1 point of is saved, and together with improved isolation another 1 point of would be saved, this choice would make the consumption of natural resources balanced (equal to cero).

Question 3

Here you see a nuclear plant. What are the advantages and disadvantages of a nuclear plant?
- Give one advantage.
- Name a disadvantage.

Advantage:

-The nuclear plants do not use too many natural resources.

Disadvantage:

- The nuclear plants decrease the score of well being and environmental scores.

Question 4

Name two strengths of this city and one weakness.

Strengths:
- The city has renewable resources to produce clean energy for its inhabitants. The city has a good energy balance.
- The city has enough light industry to keep a good economical score or status.

Weakness:
-The city does not have enough environmental areas, as wild life reserves, forests or parks.

Question 5

What is this?
How does it work (approximately)?
And what can you do with it?

- It is a Hydro plant.
- It uses the power of water to produce energy.
- You can get clean and efficient energy from this structure.

Question 6

Here, it concerns the same town as in the question 4, although one level higher. In this regard, what stands out for you/draws your attention? And, explain why you think the player made these choices?

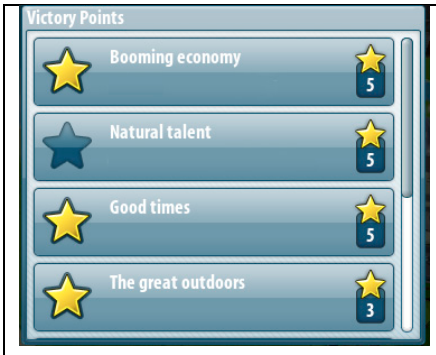
Commercial buildings replaced the light industry buildings; there are more wildlife reserves and additional solar plants.
The player may have decided to build more remunerable economic buildings, like upgrade or improve the light industry by commercial industry; the new commercial buildings need just 2 points more of environmental score and natural resources but they provide 4 times more economic benefits.
The player may have decided to build additional solar plants to have more energy.
The player may have decided to build wildlife reserves to have more well-being points and happy citizens.

Question 7

How can you improve the score of this icon in the game? (Except for building energy plants)

By implementing Sustainable Technology program in the City hall.
By placing rooftop windmills, solar roofs, in the buildings.

Question 8



This player did not earn victory points for **"Natural talent".**
What would you suggest him to win those points?
Name a possibility.

In order to earn victory points the player needs to reach a environmental score of 25, for this (s)he can build park, forest or wildlife reserves; additionally, (s)he can implement eco roofs, rainwater storage, bus stops, CO2 reduction plan, capture CO2 emissions, cradle to cradle program, recycling facilities, subway station, among other programs in the buildings.

Question 9

Consider the game you played earlier. In what way can a super windmill provide a higher amount of energy than a normal windmill?

The super windmills use vertical blades that can work with winds blowing from every direction, without the need of direction change like horizontal turbines work.

Question 10

How does time affect the score? And why?
As time pass by, each year is 1 point less in the total score. This is because if you act as soon as possible to reduce the consumption of natural resources, then there would be less impact at long term in the environment.

Appendix C: Test B with its feedback

Name: _____ Age: _____

Game Score: _____ At level: _____

Question 1

What are the three main factors influencing the total score of the game?
1. Economy score 2. Environmental score 3. Wellbeing score

Question 2

When you click on an **Economic building** of light Industry (a small business park) the menu as depicted on the side will appear:

This player has already chosen to place **recycle installations** by which he will gain 1 point for environmental score: , 2 points for natural resources: , and he will be deduced 1 point for energy balance: .

He wants to improve this business park further and therefore has to choose between the four remaining options. Select the option that you consider best.

e)

f) THIS IS THE CORRECT ANSWER

g)

h)

Why would you choose this option?

The reserves of natural resources ( ) are the only resource that cannot be "regenerated" or increased by building any structure. Therefore natural resources may be used carefully and wisely: A small business park uses 5 points of natural resources, with the recycling facilities 2 points of are being saved, and together with improved isolation another 1 point of would be saved, this choice would make the consumption of natural resources with less impact instead of spending 5 points, only 2 points of the valuables natural resources would be spend in this building.

Question 3

Here you see a coal plant small. What are the advantages and disadvantages of a coal plant? - Give one advantage. - Name a disadvantage.	
Advantage: - The coal plant small is one of the cheapest ways to produce electricity for the city. It only costs 8 “coins”. Disadvantage: - The coal plant small pollutes the air with the burning of coal.	

Question 4

	What is wrong in this game? What steps would you take to solve this problem? Name at least two steps.
There are only economy-industry buildings (business parks and industrial districts) and suburban buildings; the city does not have environmental spaces neither wellbeing buildings. Steps: 1. Build an environmental space (parks, forests, or wildlife reserves). 2. Build a wellbeing space (market, public services or stadiums).	

Question 5

	What is this? How does it work (approximately)? And what can you do with it?
- It is a Super Solar plant. - It concentrates the sunrays into the middle to produce more energy. - You can get clean and efficient energy from this building.	

Question 6

	In this situation it would be beneficial to build a Subway Station? Explain why it would be beneficial and why not?
YES: Yes, it would help to transport the citizens in a more efficient way by reducing the consumption of natural resources and help the environment in the urban districts. NO: The best at this stage is to choose an option more cheap and with electricity savings, because the electricity score is low.	

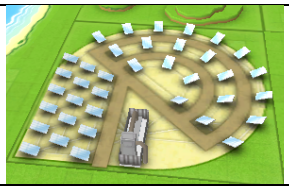
Question 7

	How can you improve the score of this icon in the game?
You cannot improve the score of natural resources, you can only maintain a low consumption of it in order to make it last longer.	

Question 8

	This player did not earn victory points for **"Good times".** What would you suggest him to win those points? Name a possibility. In order to earn victory points the player needs to reach a well-being score of 25, for this (s)he can build markets, public services, stadiums, forest or wildlife reserves; additionally, (s)he can implement cradle to cradle in the industry buildings.

Question 9

	Consider the game you played earlier. In what way can this solar plant provide a higher amount of energy than a traditional solar plant? By absorbing sunrays from different angles so it can convert into a higher amount of energy.

Question 10

When you start the game, how can you reach the level 2 as fast as possible?
The fastest way to reach level 2 is to build suburban buildings and reach the 15 inhabitants.

Appendix D: Rubrics for test A

Q	Evaluation	3 point	2 point	1 point
1	Energy sources do not need to be described precisely as in the example, but learners may have written: Nuclear fusion energy Super solar Solar energy Nuclear among others.	All 3 sources are described: 1. Nuclear fusion plant (100) 2. Super Solar (50) 3. Nuclear Plant (40)	At least 2 sources are described.	At least 1 source is described.
2	Possible explanations of why choosing the **D** option (improved insulation): - The natural resources ( ) cannot be regenerated - This option would represent the highest saving of natural resources.	Not applicable	The option D was selected. The answer includes at least 1 explanation from the evaluation section.	The option D was selected.
3	The (dis)advantages do not need to be described precisely as in the example, but learners may have written: **Advantages** Energy Electricity Less consumption of fossil fuels/natural resources **Disadvantages** Radio active / Bio hazard/Uranium Affects people, well-being, happiness among others.	Not applicable	At least 1 advantage and 1 disadvantage are described.	At least 1 advantage or disadvantage is described.
4	Possible strengths: - Renewable resources to produce clean energy - Economical buildings (light industry) - Positive wellbeing score - Good energy balance - Good economical score Possible weaknesses: - Few environmental areas (wild life reserves, forests or parks). - Bad environmental score	Identification of at least: - 2 strengths and - 1 weakness from the evaluation section.	Identification of at least: - 1 strength and 1 weakness or - 2 strengths and any weaknesses or - 2 weaknesses and any strenght from the evaluation section.	Identification of at least: - 1 strength or - 1 weakness
5	1) It is a Hydro plant. 2) It uses the power of water to produce energy. - You can get clean and efficient energy from this structure. 3a) You can get clean and efficient energy from this building. 3b) It does not contaminate / pollute. 3c) It does not consume natural resources	All 3 points are described.	At least 2 points are described.	At least 1 point is described.
6	Draw attention (difference from before and now): - Commercial buildings replaced the light industry buildings (upgrade of buldings) - Hydroplant - There are more wildlife reserves - Additional solar plants. Possible explanations: - The commercial buildings are more cost-efficient than the previous buidlings (light industry) - Increase of environmental points - Build more solar and hydro plants to have more energy (more renewable energy) - Build wildlife reserves to have more wellbeing points	Not applicable	Identification of at least: - One difference and 1 explanation - 2 explanations and any difference from the evaluation section.	Identification of at least: - One difference or - 1 explanation
7	Possible improvements: - Implementing Sustainable Technology program in the City hall. - Placing rooftop windmills, solar roofs, in the buildings.	Not applicable	Not applicable	The answer match with any option from the evaluation section.
8	To earn victory points for “Natural talent”: - Reach environmental score of 25	Not applicable	Not applicable	The answer match with any option from the evaluation section.

	- Build environmental structures (park, forest or wildlife reserves)			
9	A super windmill has vertical blades (bigger blades) that can work with winds blowing from every direction	Not applicable	Not applicable	The answer match with the description from the evaluation section.
10	Time affects score for 1 year pass is 1 point less in total score: - Act soon to have less impact at long term in environment - Every year pass by there is more consumption of fossil fuels and more savings of selected options Also as 1 year pass by more money is added to the cash reserve	Not applicable	Not applicable	The answer match with the description from the evaluation section.

Appendix E: Rubrics for test B

Q	Evaluation	3 point	2 point	1 point
1	Factors do not need to be described precisely as in the example, but learners may have written: Happiness Cash / Money / wealth /business /incoming Renewable energy /ecology Trees/nature/sustainability/CO2 emission among others.	All 3 factors are described: 1. Economy score 2. Environmental score 3. Wellbeing score	At least 2 factors are described.	At least 1 factor is described.
2	Possible explanations of why choosing the **B** option (improved insulation): - The natural resources ( ) cannot be regenerated - This option would represent the highest saving of natural resources.	Not applicable	The option B was selected. The answer includes at least 1 explanation from the evaluation section.	The option B was selected.
3	The (dis)advantages do not need to be described precisely as in the example, but learners may have written: **Advantages** Cheap Energy Electricity **Disadvantages** Pollution / Contamination Bad for the environment High consumption of natural resources among others.	Not applicable	At least 1 advantage and 1 disadvantage are described.	At least 1 advantage or disadvantage is described.
4	What is wrong in the game is one of these options: 1a) The negative environmental score 1b) Solely presence of economic, residential and energy structures 2a) The poor well-being score 2b) There are not environmental or well-being buildings (parks, forests, wildlife, market, public services or stadiums). Possible steps to improve: 1) Build / upgrade environmental structures (parks, forests, or wildlife reserves). 2) Build / upgrade well-being structures (market, public services or stadiums). 3) Change the coal plant for amore friendly renewable energy 4) Reduce CO2 emissions in city hall (taxes) 5) Add features to building to save natural resources (eco roofs, insulations, etc)	Identification of at least: - 1 factor of what is wrong and 2 steps	Identification of at least: - 1 factor and 1 step or - 2 steps and any factors or - 2 factors and any step from the evaluation section.	Identification of at least: - 1 factor or - 1 step
5	1) It is a Super Solar plant. 2) It concentrates the sunrays into the middle to produce more energy. 3a) You can get clean and efficient energy from this building. 3b) It does not contaminate / pollute. 3c) It does not consume natural resources	All 3 points are described.	At least 2 points are described.	At least 1 point is described.
6	Benefits of subway: - Reduce the consumption of natural resources - Friendly environment in the urban districts. Inconvenient of subway: - It requires 1 point of energy. - It's high cost for the economical score	Not applicable	At least 2 points are described.	At least 1 point is described.
7	The score of natural resources cannot be improved.	Not applicable	Not applicable	The answer match with the description from the evaluation section.
8	To earn victory points for "Good times": - Reach well-being score of 25	Not applicable	Not applicable	The answer match with any option from the evaluation section.

	- Build wellbeing structures (markets, public services as hospitals, stadiums)			
9	A solar plant can provide a higher amount of energy by moving the platform with the sun (absorbing sunrays from different angles)	Not applicable	Not applicable	The answer match with the description from the evaluation section.
10	A way does not need to be described precisely as in the example, but learners may have written: - Building suburban buildings to reach 15 inhabitants Increase the number of inhabitants / people - Build houses / residential area	Not applicable	Not applicable	The answer match with the description from the evaluation section.