Learning Economics Concepts through Game-Play: An Experiment*

Loretti I. Dobrescu†, Ben Greiner‡, and Alberto Motta§

Abstract

This paper introduces a video-game designed to support teaching introductory economics at undergraduate level. In order to test its effectiveness compared to traditional textbook learning we designed a laboratory experiment. Results show no evidence that playing the video-game leads to lower exam performance than reading a textbook, neither for multiple-choice nor for essay questions. We also find no gender bias and no effect of announcing the test prior to the learning task or thereafter. However, game behavior appears to be related to test performance, and differently so for different types of learning. Students perceive the two learning tools similarly in terms of understanding requirements or usefulness, but enjoyed the video-game considerably more. Interestingly, although women enjoyed the game less than men, they do not differ in their test performance.

Keywords: economics education, new teaching methods, video-game, experiment.

JEL Classification: C72, C92, D3

*We thank Loretta O'Donnell, Kevin Fox, Stephen Marshall and Prem Ramburuth for continuous support and encouragement. We are also grateful to Dror Ben-Naim, Danny Carroll, Gigi Foster, Peter McGuinn and Malcolm Ryan for discussions and comments. Financial support from an Australian School of Business Research Grant and the IT Investment Plan (ITIP) fund at the University of New South Wales is gratefully acknowledged. We further thank Martin Slater and James Murchison for outstanding support with the software design and in conducting the experiments.

†University of New South Wales, School of Economics, UNSW Sydney NSW 2052, Australia, Tel: 61 2 938 57480, e-mail: dobrescu at unsw.edu.au.

‡University of New South Wales, School of Economics, UNSW Sydney NSW 2052, Australia, Tel: 61 2 938 59701, e-mail: bgreiner at unsw.edu.au.

§University of New South Wales, School of Economics, UNSW Sydney NSW 2052, Australia, Tel: 61 2 938 59771, e-mail: motta at unsw.edu.au.

1 Introduction

In the last decade, the spread of information technology in education has been rapid. The extensive use of teleconferencing, e-learning, and other advancements involving telecommunication technologies has considerably influenced the educational process in ways that complement and, sometimes, supersede traditional methods like face-to-face lectures (Bennett, Mims and McKenzie, 2002; Donovan, Figlio and Rush, 2006; Figlio, Rush and Yu, 2010; Kearsley, 2000; Schreiber and Berge, 1998). More recently, complex "serious games" have emerged, significantly changing our perception of learning and approach to teaching.

Serious games are basically simulation games in which the players can experiment with different decisions and analyze their consequences in a virtual environment. While currently there is an accelerating tendency to use these games for teaching purposes, clean evidence on the effectiveness of this new learning environment is rare (see Sitzmann, 2011, for a meta-analytic study of the instructional effectiveness of computer-based simulation games). We aim to fill this gap by reporting results from a laboratory experiment in which students either read one chapter of a microeconomics textbook or played a level of an educational video-game. The latter, via game play and short-videos, conveyed the exact same information as the textbook. Thus, the only difference between these two groups lies in how the material was delivered: Some students read the textbook chapter, while others played the video-game.

After reading the chapter or playing the game, students took part in an exam, comprised of multiple-choice questions and free-text essay questions. The performance in the exam was monetarily incentivized. In addition, we administered a questionnaire that collected various demographics and measures of learning satisfaction. As a result, we are able to assess: i) the students' cognitive learning outcomes, using the score they achieved by answering multiple-choice and essay questions, and ii) the students' affective (or behavioral) learning outcomes, using the information on their satisfaction with this learning tool and their perception of its helpfulness. By analyzing these outcomes, we are able to determine how the interactive video-game delivery of the lecture compares with the passive textbook delivery across both learning dimensions. Additionally, in order to identify which students can be expected to benefit more from educational video-games, we consider two incentive conditions. In one condition ("high-powered incentives"), participants were informed about the exam-like test and its monetary incentivization right from the beginning of the session, so they could immediately adapt their behavior. In a second condition ("low-powered incentives"), participants were only informed that there would be a second part of the experiment, and to expect instructions on this part only after they finished the first part. While they might have had guesses about what was going to happen in the second part, their incentives to take the first (learning) part seriously should be lower than in the high-powered condition.

In addition to cognitive and affective outcomes, a special software allowed us to track every action undertaken by the players in the game, ranging from clicking on the screen to replaying a video, or hovering over a graph. This level of detail gives us the opportunity to study in-game behavior, and its connection to cognitive and affective outcomes.

We can report several interesting results. First, playing the video-game does not lead to lower exam performance than reading a textbook chapter. This holds true for multiple-choice questions as well as for essay questions. Second, the strength of incentives (announcing the exam or not) seems not to have an effect on final scores in either environment. Third, we do not find evidence that playing the video-game would have a different relative effect on exam performance for males compared to females. Fourth, contrary to previous findings on business simulations (e.g. Anderson and Lawton, 1992; Gosen and Washbush, 2004; Washbush and Gosen, 2001), we find that game behavior and exam performance are related, and differently so for different types of learning and assessment: scores in multiple-choice questions are only correlated with game play elements; essay question scores are correlated with both watching instructional videos and engaging in game play; and, particularly for questions requiring deep understanding, synthesis and evaluation, using interactive graphs is also related to exam performance. Finally, students perceived the two learning tools similarly in terms of understanding instructions or usefulness, but as expected, they enjoyed the video-game considerably more. Interestingly, although women enjoyed the game less than men, they performed at the same level in both the multiple-choice and the essay questions of the test.

It is worth noting that our results on the differential enjoyment in the game vis-a-vis the textbook may have further implications that could not be captured in our experimental environment where the allotted time was fixed and the treatment conditions were exogenously assigned (rather than endogenously self-selected by the students). A rational model of learning behavior would predict that students optimally allocate more time to studying when the available learning material is a more enjoyable game rather than a less enjoyable textbook. The results from our experiment then might imply that playing the game could lead to higher exam performance than reading a textbook, provided that the environment is one in which students can allocate their time freely across different activities.

Our contribution is related to a number of papers in the field of education inquiry. We discuss these papers in the theoretical background section and when formulating our hypotheses for the laboratory experiment. A few general notes, however, are in order here. Currently, there seems to be little empirical evidence on the impact of educational video-games on students' educational outcomes. An extensive base of publications exists on computer-based business simulations. However, even in this field, there is almost no evidence from cleanly designed experiments that compare learning outcomes under alternative pedagogies (two exceptions are Chou, 2012, and Figlio et al., 2010, but they focus on online learning tools and not simulations). Wolfe (1990) and Burns, Gentry and Wolfe (1990) identified this problem already 20 years ago, but there is still no convincing evidence on what simulations accomplish regarding cognitive learning. Gosen and Washbush (2004) conduct an analysis of 115 studies on this topic and find "that none [...] met all the criteria for sound research" (p. 283). Similarly, Wideman, Owston, Brown, Kushniruk, Ho and Pitts (2007) conclude that support for the educational impact of simulations is subjective, at best.

A similar problem exists when analyzing the attitudes of the participants exposed to simulations (i.e., the affective domain). Most of the related studies, with few exceptions, find that students like simulation exercises and view them more positively than either lectures or case discussions (Burns et al., 1990; Faria, 2001; Gosen and Washbush, 2004). However, these are relative comparisons made by students experiencing different learning tools within a course. Anderson and Lawton (2009) note that providing valid substantiation for the power of a learning tool requires rigorous experimental design and exogenous treatment assignment. To the best of our knowledge, no study so far has employed an experimental design featuring control groups and allowing comparisons between student attitudes in a course that is solely textbook based versus those in a class that is solely game based. Our contribution attempts to fill this gap by combining experimental design with a rich data set of in-game behavior.

From a policy perspective, more evidence on the effectiveness of educational video-games is needed. If these games were reasonable, cost-effective substitutes for traditional classroom lectures, delivering high-quality content, then they could help address the problems of increased live-lecture class sizes (Bettinger and Long, 2007) and financial constraints in higher education (Figlio et al., 2010). Additionally, given their interactive and goal-oriented nature, it might be that these games can help overcome students' weak incentives to regularly keep up with classes, and prevent last-minute cramming in online courses (Donovan et al., 2006). Given their promise, a sound testing of their effectiveness in supporting learning is vital. This evidence would also further support our understanding of which students can be expected to benefit more from (or be hurt by) these new learning tools. Such questions are crucial for university educators seeking to understand the impact of their course design decisions on students perceptions and behavior, and eventually on their academic outcomes.

The remainder of the paper is organized as follows: We first introduce the economic video-game used in this study, and describe the level played by the students in the experiment. We then turn to the theoretical background and lay out the hypotheses for our experiment. Next, we present the methodology and show our experimental design and procedures. In the last two sections we discuss our results and conclude, respectively.

2 The Economic Video-Game

The main video-game covers roughly the material offered in a standard Microeconomics I introductory undergraduate course. The content is delivered by placing the player/student into an imaginary world populated by various agents (our online supplementary materials[1] include a demonstration video of the level described below, as well as game screenshots). Each level consists of a map, which features a certain number of places, goods and agents. The player controls only one of these agents. If more than one agent populates a map, the computer controls the remaining agents ('non-playing characters', hereafter NPCs). The player can move her agent to a certain point in the map by clicking on that point. If the player clicks on a specific location, the agent moves to that place and, upon arrival, interacts with it. Each place allows the player to acquire a certain amount of a given resource. The game consists of a number of levels and sub-levels, which correspond to textbook chapters. Depending on the specific level, the player can use the acquired resources in different ways: resources can be consumed, used to build a house of the player's choosing, traded, and so on. In a later level, the player is asked to use her resources to build a firm and, subsequently, she is asked to make production decisions.

In our experiment, we only use a limited section of the game. This section covers the material presented in Chapter 2 of Frank, Jennings and Bernanke (2012), titled 'Comparative Advantage: The Basis for Trade'. Past experiences with Microeconomics I students at the University of New South Wales (hereafter UNSW) suggest that this chapter is particularly difficult to comprehend, as reflected by the corresponding learning outcomes. The material requires familiarity with a rich array of economic tools ranging from diagrams to basic math, and establishes the important economic concepts of comparative and absolute advantage (see Sloman (2002) non-computer-based game facilitating the comprehension of this topic).

In our video-game, the respective section used in the experiment contains six sub-levels, plus a tutorial. The player is required to master the levels in a certain order, but upon

[1] http://ben.orsee.org/supplements

completion of a level, she is allowed to replay it at will, in any order and at any time. The game is fully self-explanatory – all the instructions necessary to play the game and use it as a learning tool are offered in a tutorial, which the player needs to successfully complete before being allowed to play.

The game uses the same numerical examples as Chapter 2 of Frank et al. (2012). In describing the concept of the Production Possibility Curve (hereafter PPC), for example, the textbook offers examples of agents with different productivities and, consequently, different PPCs. The game provides the same kind of information, but using a very different approach.

First, the player selects one production location by clicking on it. A standard graph with one good on the x-axis and the other on the y-axis is displayed in the top-right corner of the screen. This graph is updated in real time to depict the combination of the two goods the player has produced until that point in time. Such a standard figure appears in every introductory microeconomics textbook, but in the video-game the player can shape the graph with his/her own actions and observe the consequences in real time. By playing this level the student realizes that certain combinations of goods are attainable, while others are not (i.e. he/she discovers the PPC).

In the second part of this level, another agent is introduced. This agent is controlled by the computer (NPC) and the player can offer him to exchange a certain amount of one good for another. In this process, the player learns the basics of the bargaining process, as the offer must be low enough to be acceptable to the NPC but high enough to be profitable. In order to make a profitable deal, the player must also specialize according to his/her "comparative advantage": he/she should specialize in the productive action he/she is relatively more efficient at. By offering the right amount and specializing in the right activity, the player can end up consuming a combination of goods that was not attainable in the absence of trade - a point outside the PPC.

There are other analogies between the game and the textbook. The former presents the player with primary objectives that need to be completed to access the next levels. These correspond to the main numerical examples in the textbook. Secondary objectives are also offered in the game, but they are entirely optional and they make use of the numerical examples that the textbook offers as side-exercises in the main text.

Aside from game play, the game features a media library that can be accessed by the player at any point in time. The media library displays the definitions of the basic economic concepts. To draw a parallel with the textbook, these are the concepts that are highlighted at the margin of each page in the textbook chapter. The media library also provides access to videos that were developed specifically for the game. These videos consist of voice recording plus screen annotations. Most of the concepts were conveyed via game play, with some crucial concepts also reinforced by the videos.

3 Theoretical Background and Hypothesis Development

Part game and part simulation, the economic video-game falls nicely into the definition of simulation games proposed by Sitzmann (2011): "computer-based simulation games [are] instructions delivered via personal computer that immerses trainees in a decision-making exercise in an artificial environment in order to learn the consequences of their decisions." There are several theories that can help evaluate the effectiveness of simulation games. The earliest ones emphasize their motivational potential: Intrinsically motivated players exert more effort, enjoy learning more and ultimately experience deeper learning altogether (Malone, 1981). A well-designed simulation game has also the potential to induce a positive cycle that creates a "flow state" where players experience total immersion in the simulation (Csikszentmihalyi, 1990; Garris, Ahlers and Driskell, 1981). More recently, Tennyson and Jorczak (2008) propose that simulation games should be more effective than other instructional methods, not only because they enhance motivation (i.e., the player's affective structure), but also because they simultaneously target other elements such as memory and knowledge base (i.e., the player's cognitive structure). And since the cognitive and affective structures interact together, they can be targeted simultaneously to enhance learning outcomes via transformational play (Barab, Gresalfi and Ingram-Goble, 2010).

All these theories suggest that simulation games have the potential to enhance learning outcomes (see also Squire, 2006) through a combination of entertainment and active learning (Bell and Kozlowski, 2008; Frese, Brodbeck, Heinbokel, Mooser, Schleiffenbaum and Thiemann, 1991). Entertainment keeps the players engaged, while active learning induces them to exert more cognitive effort to process and retain information (Bell and Kozlowski, 2008; Brown and Ford, 2002; Jonassen, 2002). Finally, simulation games also have the potential to increase users' satisfaction with the learning experience (Sitzmann, Brown, Casper, Ely and Zimmerman, 2008).

A variant of simulations are in-class experiments, which are frequently part of the teaching strategy of introductory microeconomics courses (Emerson and Taylor, 2007) and also aim at triggering active learning and retention through engagement. While several studies documented the positive impact of this teaching tool on student performance (Dickie, 2006; Emerson and Taylor, 2004), it appears that its effects are limited as far as choosing an Economics major is concerned (Emerson and Taylor, 2010).

Simulation games and variants may however involve considerable drawbacks, for instance the substantial time commitment they require, or their inability to effectively teach terminology and basic concepts (Faria and Wellington, 2004; Fripp, 1993; Lean, Moizer, Towler and Abbey, 2006; Saunders, 1997). Gender bias is another concern: females were found to report more computer anxiety, less computer self-efficacy, and less favorable computer attitudes (Chirieac, Burns and Case, 2000; Jackson and Ervin, 2001; Jazwinski, 2001). They appear to be less interested in digital games, play less frequently (Lucas and Sherry, 2004; Wright, Huston, Vandewater, Bickham, Scantlin, Kotler, Caplovitz and Lee, 2001), and perceive computer-aided learning as generally less useful than men (Venkatesh and Morris, 2000).

In the following, we rely on the theoretical framework presented above to develop a number of hypotheses, which are then tested using a laboratory experiment.

Hypothesis 1 *Game play leads to lower exam performance than reading a textbook chapter.*

The theoretical literature suggests that simulation games have the potential to outperform more traditional instructional methods by enhancing motivation, effort, and reactions (Garris et al., 1981; Malone, 1981; Sitzmann et al., 2008; Tennyson and Jorczak, 2008). But the success of games as a learning tool critically depends on the learning content and environment. The empirical literature, for instance, suggests that business simulations are an inefficient method for teaching terminology, factual knowledge, basic concepts and principles when compared to face-to-face lectures (Faria and Wellington, 2004; Fripp, 1993; Lean et al., 2006; Saunders, 1997). Further, simulations and games require a substantial time commitment to be effective, and are particularly helpful when participants are free to engage with the learning tool at will (Sitzmann, 2011). Finally, game play is likely to be more effective a learning tool for some students than others, and in an unrestricted setting we would expect students to gravitate toward the resources most suited to them.

Our experiment includes each of these adverse effects: We use an educational game to teach the (abstract) economic concept of comparative and absolute advantage in a limited amount of time (students in both conditions had one hour to learn, including the approx. 5 minutes tutorial in the game condition), with exogenous assignment to conditions. Hence, given the findings discussed above, we expect that those who play the video-game in our experiment will perform worse in the exam than those who read the textbook chapter.

Hypothesis 2 *The relative exam performance after game play compared to textbook reading is better when incentives are low-powered.*

In order to capture the differential effect of the game and textbook on students with different degrees of motivation, we examined two different incentive conditions. With high-powered incentives, participants were informed about the test and its payoff rules right from the start. In the low-powered incentives condition, participants were informed about the second part of the experiment only after they finished the first part (see Section 4 for details). As a result, without information on what it would be useful for, the participants in the low-powered incentives condition might have been less motivated to learn the material. Our conjecture is that the underperformance of game play compared to textbook participants is lower for the group of low-powered incentives students. One obvious reason is that the game could be fun to play and entertaining in itself (Csikszentmihalyi, 1990; Garris et al., 1981). Students that lack intrinsic motivation might find it easier to focus on the game because it is more interactive and stimulating than a textbook (Bell and Kozlowski, 2008; Brown and Ford, 2002; Jonassen, 2002). The game also provides feedback with student-specific comments about errors and suggestions for improvement. This can be more beneficial for low-motivated students than for highly motivated ones who are committed to improve their performance and therefore independently assess their own skills. Another difference between the game and the textbook is that the game encourages students to focus their attention on the task rather than on simply getting the right answer and a higher mark (Bangert-Drowns, Kulik, Kulik and Morgan, 1991; Shute, 2008). This could induce unmotivated students that believe they lack some innate abilities to exert an additional amount of effort as opposed to reinforcing the notion that they are "doomed" (Hoska, 1993).

Hypothesis 3 *The relative exam performance after game play compared to textbook reading is better in essay questions than in multiple-choice questions.*

We use two measures of cognitive outcomes, namely the students' score in multiple-choice questions and in essay questions of an incentivized exam conducted after the learning task ended. These two instruments are the most prominent in assessing learning in economics, and have also been widely used to measure the learning effectiveness of simulation exercises (Washbush and Gosen (2001)). Our hypothesis that performance effects are different for essay questions and multiple-choice questions is related to the level of knowledge that the two tasks (i.e., playing the game or reading the book) induce. Sugrue (1995) identifies three levels of knowledge: 1) understanding concepts; 2) understanding principles that link concepts; 3) understanding the application of the concepts and principles under the appropriate circumstances. Essay questions require an elaborated written answer for which students have to use their knowledge in a new context. Thus, in Sugrue (1995)'s taxonomy, essay questions belong to the third level of knowledge. The multiple-choice questions on the other hand, are generally focused on reproduction, and so they assess the first level of the knowledge structure, namely understanding terminology and basic concepts, something that simulation games could be ill-equipped to teach (Faria and Wellington, 2004; Fripp, 1993; Lean et al., 2006; Saunders, 1997). But, owing to the interactive elements and the possibility to experiment with different scenarios, games might be better at illustrating the dynamics of a problem. As a result, they train the player to apply concepts and principles to different external conditions, which is exactly what essay questions test.

Hypothesis 4 *Male students learn better through games than female students.*

There are several mechanisms through which gender differences may be operating in our experiment, all of which suggest a better learning and performance effect of game play for males. First, gender has been an important factor in explaining the adoption of computers and online technologies. In the early stages of technology penetration, researchers believed that this area was male dominated (Gregory, 1997). Since then, several studies on this issue have investigated gender differences (Chirieac et al., 2000; Jackson and Ervin, 2001; Jazwinski, 2001) and generally found that females reported more computer anxiety, less computer self-efficacy, and less favorable computer attitudes. Women's rating of perceived usefulness of computer-aided learning is lower than men's (Venkatesh and Morris, 2000). Females have less interest in digital games, less game-related knowledge, and play less frequently (Lucas and Sherry, 2004; Wright et al., 2001), which in turn will favor males. Second, male students might be more tempted to defer learning and cram for the exam, if procrastination is possible. Being strongly attracted to action-oriented competitive games (Vorderer, Hartmann and Klimmt, 2006), it might be that an interactive educational game will help them stay engaged with the material. And third, women may also be averse to the competitive nature of game playing. This would be in line with the experimental economics finding that women are significantly less competitive (and perform worse in competitive environments) than men (Croson and Gneezy, 2009; Gneezy, Niederle and Rustichini, 2003; Niederle and Vesterlund, 2007, 2011).

Hypothesis 5 *Game behavior and exam performance are related; participants who are more active and score higher in the game also perform better in the exam.*

Players can be more active in a video-game because they are attracted to the challenge of 'winning' the game, or because they simply enjoy playing the game, regardless of the outcome (i.e., the score) they achieve. In any case, higher involvement with the game's tasks implies more active learning, and so it may be a predictor of performance in the learning assessment (Bell and Kozlowski, 2008; Frese et al., 1991). The research on business simulation, however, has not found a link between simulation performance or behavior and learning (Anderson and Lawton, 1992; Washbush and Gosen, 2001). Gosen and Washbush (2004) conducted an extensive review of the existing research to assess the effectiveness of business simulations on learning. They concluded that performance on a simulation exercise should not be used as a proxy for learning. However, our starting hypothesis is the intuitive link between activity and learning.

Hypothesis 6 *Students who enjoy themselves more when playing the game score higher in the exam.*

Our theoretical frameworks suggest that positive emotional arousal and enjoyment should have positive effects on self-efficacy (Csikszentmihalyi, 1990; Garris et al., 1981; Malone, 1981; Tennyson and Jorczak, 2008). The literature on affective learning (see Faria, 2001) found support for a positive relation between participants' attitude toward the simulation and the simulation's role in their perceived learning. But Faria (2001) also points out that affective learning is, by its inherent nature, based on the perceptions of the simulation's participants, not objective assessments. Our experiment allows us to objectively assess whether positive attitudes indeed correlate with a higher performance in the test. We caution, however, that such correlation does not necessarily imply causation, since it could also be that better-performing students find the learning experience more enjoyable. Our textbook conditions might serve as a control here.

4 Methods: Experimental Design and Procedures

The experiment was conducted in July 2012, in Week 1 of the second term, at the ASB Experimental Research Laboratory at UNSW. We chose this week as the time of the experiment because the topic tested in the experiment was then taught in Week 2. Eighty-one undergraduate students at UNSW were recruited using the recruitment system ORSEE (Greiner, 2004). To test the effectiveness of the game in different student populations, we recruited both students who were signed up for the Microeconomics I course and students from other fields who were not enlisted for this course. (This information was verified in the post experimental questionnaire.) None of the researchers involved in this project was teaching this course in that term.

When subjects arrived, they were signed in, assigned to a cubicle (a table with a computer, mouse and keyboard, and dividers on three sides so they cannot communicate with other participants), and were asked to read the participant information statement, sign the consent form, and read a first set of instructions (see Appendix B available online). Each session consisted of two parts. In the first part, lasting exactly 1 hour, participants either read the textbook chapter, or played the corresponding level of the video-game. In the second part, participants took part in a test. In all conditions, for the entire duration of the experiment session, participants were not allowed to communicate with each other or to use the Internet or their mobile phones. Participants were provided with pen and paper for making notes (which were collected before the test was taken). In the game treatments, the time allotted to play the game included time needed for instructions and a video tutorial on how to play the game. Participants interacted with the game at their own pace. A research assistant provided assistance for technical issues only.

As the subject of our experiment we chose "comparative advantage", which is the topic taught in Week 2 of the microeconomics course at UNSW. As we mentioned before, the topic requires the use of a rich array of economic tools, and establishes important economic concepts. Moreover, this part of the material is oftentimes problematic for students, as reflected by past exam scores. In our baseline treatment condition, we provided students with Chapter 2 of the corresponding textbook (Frank et al., 2012). In the game treatment condition, we used the level on comparative advantage that we described in Section 2. A copy of the textbook chapter and an executable trial version of the game are available from the authors on request. Screenshots of the game, experimental instructions, and all multiple-choice and essay questions used in the test can be found online (see Footnote 1).

In one incentive condition (high-powered incentives), participants were informed about the test and its incentivization right from the beginning of the session, and so could adapt their behavior in the first part to these incentives. In a second condition (low-powered incentives), participants were only informed that there would be a second part of the experiment, and they would receive the details on that part only after they finished the first part.

The test consisted of eight multiple-choice questions and five open-ended questions (see Appendices C and D available online). Obviously, the materials handed out in Part 1 did not include these questions. In designing the test, we tried to obtain a set of questions as objectively selected and representative as possible. The eight multiple-choice questions represented all the questions on the topic of the experiment used in the UNSW microeconomics exams in the three years before the experiment. The questions were handed out on paper, but were answered on the computer screen. For instance, we asked:

Question 3 *Given the table below showing the resource costs (hours of labour per unit of output) of producing cars and wheat in Australia and Canada, which of the following is true?*

	Canada (Labour hours)	Australia (Labour hours)
Cars	10	40
Wheat	20	30

a) Australia has an absolute advantage in producing cars; b) Australia has an absolute advantage in producing wheat; c) Canada has an absolute advantage in producing cars and wheat; d) Australia has a comparative advantage in producing cars.

Question 4 *A Canadian worker can produce four tonnes of rye in a year or produce three bales of wheat in a year. An Australian worker can produce two tonnes of rye in a year or one bale of wheat in a year. With trade, the price of rye in terms of wheat will be: a) 1 bale of wheat; b) Between 1/2 and 1 bale of wheat; c) Between 3/4 and 1/2 bale of wheat; d) Between 1 and 4 bales of wheat.*

We used the experimental standard software zTree (Fischbacher, 2007) to program the interface. The five open-ended essay questions were chosen from the end of chapter questions in the textbook. We used all four questions included in the "Review Questions" Section, as well as one question randomly selected from the "Problems" Section. Two examples are:

Question 3 *In what way are an individual's opportunity costs of producing related to the rate at which they will be exchanged or traded?*

Question 4 *Why does saying that people are poor because they do not specialize make more sense than saying that they perform their own services because they are poor?*

Only Part 1 or Part 2 were selected for payoff at the end of the experiment. If Part 1 was selected, each participant received a flat payment of $30 in addition to their show-up fee of $5. If Part 2 was selected, participants received $3 for each correctly solved multiple-choice question. In addition, each participant had a 1/10 chance to be paid for one of her answers to the five essay questions. The participant first threw a ten-sided dice to determine whether she would be paid for an essay question or not. If selected, she threw the dice again to determine exactly which question would be paid. Participants knew that the question would be marked according to typical exam standards, and that they could receive up to 25 points for a question. Each obtained point would pay them \$10. Thus, participants could earn up to \$250 in this part. Show-up fees, flat payments (if Part 1, the learning task, was selected) and multiple-choice payments (if Part 2, the exam, was selected) were paid in cash right at the end of the experiment. Payments for essay answers were paid out in cash one week after the experiment, since the answers had to be marked first.

The open-ended questions were given to two independent markers. The markers were PhD students experienced in marking undergraduate microeconomics exams, and were not informed about our research hypotheses or the treatment conditions of the exams. Both markers marked all exams and received a flat payment based on the time they spent marking. The markers could assign between 0 and 25 marks to each exam question.

Marker 2 gave significantly higher marks than marker 1 (avg. of 67.0 vs. 50.4). This is true at the aggregate level over all questions (Wilcoxon Signed Ranks Matched Pairs test, $p < 0.001$) as well as the level of each question except question 5 ($p = 0.042$ for question 1, $p < 0.001$ for questions 2, 3, and 4, $p = 0.1345$ for question 5). Nevertheless, the correlation between marks given by marker 1 and 2 was very high, with correlation coefficients of 0.70 (0.83, 0.39, 0.48, 0.50, 0.92) for the overall mark (marks for questions 1, 2, 3, 4, and 5, respectively, all p-values smaller than 0.001). Given these high and consistent correlations, we consider our marking data as reliable, despite the level shift between markers. We use the average of the marks assigned by the two markers in our analysis.

In addition to the experimental sessions described above, we conducted two other sessions in March 2014 to check whether our textbook and game conditions yielded positive learning effects. The only change in these sessions was that instead of reading the textbook chapter or playing the game, students could spend their time as they wished (including using the Internet). As in our low-powered incentive conditions, the exam was only announced after the "free time" part ended. This set of sessions, conducted in Week 2 (just before the respective Microeconomics I lecture) and later in Week 4 of the semester, yielded three different baseline subject groups: 13 students enrolled in Microeconomics I who (self-reportedly) had prepared for the week's lecture about comparative advantage, 16 students in Microeconomics I who did not prepare, and 21 students who were not currently enrolled in the course, but could take it at a later time. These comparison groups will allow us to confirm the existence and extent of learning through textbook and game play.

5 Results

Our experimental design allows us to assess both the cognitive and the affective (or behavioral) outcomes that students derive from the learning process. Traditionally, deep learning has been summarily evaluated through assessment scores. However, in a climate emphasizing student evaluations, the affective outcomes have become increasingly important. There are generally two types of affective outcomes: i) expressed satisfaction with a learning tool, and ii) specific perceptions of particular skill development.

In the following we report the results from our experiment on both cognitive and affective learning outcomes. Subsection 5.1 presents summary statistics for our subject pool. Subsection 5.2 is devoted to our main outcome variables, the marks achieved in the multiple-choice and essay-type questions, and analyzes their relation to our main treatment parameters and demographic characteristics of the participants. In Subsection 5.3 we investigate the learning effects of our treatment conditions, while Subsection 5.4 presents the relationship between in-game behavior and test outcomes. Finally, Subsection 5.5 shows our data on students' perception of the learning environments.

5.1 Subject pool

For our experiment we recruited participants from the subject pool of the Australian School of Business Research Laboratory. Both students taking the Microeconomics I course and those not in this class were eligible. Participants signed up for sessions not knowing about the experiment's content, purpose or conditions. Table 1 shows the distribution of subjects' demographic characteristics among our four treatment conditions. We obtained a mostly balanced subject pool across the treatments. For all characteristics listed in Table 1 we statistically compared participants in-game and textbook conditions, as well as participants in low- and high-incentive conditions. The only statistically significant imbalance we find is with respect to whether students are signed up in the Microeconomics I course (Fisher's exact test comparing proportions of students signed up for Micro I, p=0.045) and how they rate their familiarity with the topic (Wilcoxon Ranksum test comparing median rating on familiarity scale, p=0.042). (These two variables are also highly correlated with each other.) In our analysis below, we will run robustness checks that control for these characteristics, as well as for other demographic variables.

TABLE 1: Demographics of subjects in the four different treatments

	Textbook chapter		Video-Game		C vs. G	L vs. H
	Low Inc.	High Inc.	Low Inc.	High Inc.	p-value	p-value
N	19	20	22	20		
Age	19.9 (2.8)	19.4 (2.2)	19.5 (2.4)	19.5 (1.4)	0.707	0.872
Male	68%	85%	59%	65%	0.158	0.337
International	37%	20%	41%	45%	0.246	0.645
Field of study					0.389	0.178
Commerce & Econ	47%	23%	48%	44%		
Engineering	26%	53%	33%	50%		
Science	26%	24%	19%	6%		
HS Score	87.7 (17.3)	91.4 (8.3)	91.8 (8.3)	91.5 (7.7)	0.629	0.994
WAM	66.8 (13.4)	71.5 (7.3)	70.3 (9.5)	67.7 (12.3)	0.754	0.700
Enrolled in Micro I	58%	45%	64%	85%	0.042	0.819
Familiarity with topic (1-5)	1.9 (0.8)	1.8 (0.9)	2.2 (1.2)	2.5 (0.9)	0.045	0.804

Note: Standard deviations are given in parentheses. The last two columns list p-values from Wilcoxon Ranksum tests (on differences in medians) and Fisher's exact tests (on differences in proportions) between Chapter-Game conditions and Low-High incentive conditions, respectively. The exact statement assessing topic familiarity was "My level of knowledge on the topic before this experiment was".

5.2 Treatment effects

In Table 2, Models (1) and (3) report results from Tobit regressions of the number of multiple-choice questions solved correctly (out of 8) and the total marks achieved in essay questions (out of 125) on treatment dummies. Interestingly, we observe no treatment effects in our data at the aggregate level. Neither one of the two treatment dummies nor their interaction is significant in explaining the aggregate test results. In general, this is also true when applying the same analysis at the level of the single questions (likelihood of answering a multiple-choice question correctly, marks achieved in a particular essay question).

TABLE 2: Tobit regressions of test results on treatment conditions and gender

Dependent	Correct MC questions		Total marks in essay questions	
Model	(1)	(2)	(3)	(4)
Constant	4.58***	3.81***	55.97***	58.74***
	[0.38]	[0.57]	[4.9]	[7.45]
Game	-0.51	0.23	0.89	-4.8
	[0.52]	[0.73]	[6.68]	[9.59]
High Incentive	0.10	-0.09	2.95	3.62
	[0.53]	[0.53]	[6.84]	[6.94]
Game x High Incentive	-0.32	-0.13	3.26	2.3
	[0.73]	[0.73]	[9.5]	[9.56]
Male		1.13*		-4.04
		[0.63]		[8.23]
Male x Game		-1.07		8.98
		[0.81]		[10.65]
N	81	81	81	81
Left-(non-)right-censored	1(79)1	1(79)1	0(81)0	0(81)0
LogLL	-154.8	-153.2	-362.8	-362.4

Note: Standard errors are given in brackets. *,**, and *** denote significance at the 10%, 5%, and 1% level, respectively.

To test Hypothesis 4 on the effectiveness of the video-game for different genders, we re-run these regressions and include a dummy for the gender of the participant, as well as its interaction with our game condition. We find a weakly significant general effect of being male only on the number of correctly solved multiple-choice questions (not for essay scores), but no evidence that playing the video-game would increase the relative performance of males compared to women. The interaction coefficient has opposite signs for multiple-choice and essay questions, and is insignificant in both cases. The general effect of being male for the multiple-choice questions score disappears however when we drop the interaction effect, and remains insignificant when adding further control variables (see Table 3 below). Thus we do not interpret this effect as a robust result.

TABLE 3: Tobit regressions to test robustness of main results

Dependent	Correct MC questions			Total marks in essay questions		
Model	(1)	(2)	(3)	(4)	(5)	(6)
Constant	4.42***	4.30**	3.98	52.14***	87.58***	50.43
	[0.44]	[1.81]	[2.10]	[5.67]	[23.64]	[27.06]
Game	-0.52	-0.44	-0.65	0.51	0.32	0.02
	[0.52]	[0.51]	[0.51]	[6.62]	[6.64]	[6.56]
High Incentive	0.13	0.08	-0.14	3.81	1.42	-0.26
	[0.53]	[0.53]	[0.53]	[6.80]	[6.94]	[6.88]
Game x High Incentive	-0.41	-0.30	-0.05	0.99	4.52	7.75
	[0.75]	[0.73]	[0.73]	[9.56]	[9.47]	[9.37]
Enrolled in M1	0.27			6.62		
	[0.40]			[5.09]		
Gender		0.61	0.75*		2.05	2.93
		[0.41]	[0.41]		[5.35]	[5.46]
Age		-0.01	-0.04		-1.64	-1.61
		[0.09]	[0.09]		[1.21]	[1.18]
International		-0.31	-0.21		0.35	-3.40
		[0.43]	[0.44]		[5.61]	[5.63]
Commerce/Economics		0.67*	0.68		-0.98	0.15
		[0.39]	[0.39]		[5.08]	[5.01]
WAM			0.01			0.53**
			[0.02]			[0.23]
N	81	81	80	81	81	80
Left-(non-)right-censored	1(79)1	1(78)1	1(56)1	0(81)0	0(81)0	0(80)0
LogLL	-154.5	-152.9	-104.3	-362.0	-361.7	-354.4

Note: Standard errors are given in brackets. *,**, and *** denote significance at the 10%, 5%, and 1% level, respectively.

To test the overall robustness of these results we re-run our specifications including demographics and other controls. Table 3 reports the results. In Models (1) and (4), besides our treatment dummies, we also include a dummy indicating whether the student was signed up for the Microeconomics I course or not. Being signed-up has no effect on the test outcomes in terms of correct multiple-choice questions or essay questions marks in either model, nor does its inclusion change the estimates and significance of the treatment parameters.

Models (2) and (5) include controls for demographics. In particular, we included age, gender, whether the student is international, and whether the student studies economics, commerce or a related field as explanatory variables in the analysis. Demographics seem not to play a role for test results, and our Null result of no treatment effects is robust against the inclusion of these variables.

Finally, in Models (3) and (6) we also include the student's previous weight-point average (WAM score) from the university's records (with one observation missing, and 23 WAMs calculated from concurrent courses, excluding Microeconomics I, due to no previous marks received). As one would expect, the WAM score is positively correlated with the number of correct multiple-choice questions and the marks achieved in essay questions, but this is only significant for the latter. Interestingly, including the WAM score in the model on multiple-choice questions yields a significant positive coefficient on the dummy on gender. This indicates some potential multicollinearity across these two variables.

In sum, we do not find support for Hypothesis 1. In particular, we cannot reject statistically the Null hypothesis of the same test performance after playing the game for one hour *vs.* reading the textbook chapter within that hour. In addition, there is no evidence of incentive environment effects (Hypothesis 2) or differential learning environment effects for performance in multiple choice and essay questions (Hypothesis 3). We also find no evidence that males' performance reacts more strongly to game environment (neither at the aggregate nor at the individual question level), thereby not lending support to Hypothesis 4.

5.3 Learning effects

To confirm that both textbook and game play lead to a positive learning effects (i.e., to ensure that they both are not *equally ineffective*), we ran two additional control sessions. The new sample included both Microeconomics I students who prepared for the comparative advantage lecture and who did not, as well as students who were not enrolled in Microeconomics I, but could take it at a later time. Additionally, by asking students to assess their prior knowledge on comparative advantage in the post-experimental questionnaire, we classified them into those with poor prior knowledge (having answered "poor" or "very poor" to the question) and good prior knowledge (answers "ok", "good", or "excellent"). If the learning methods were effective, we would expect that students with poor prior knowledge achieve higher test scores when they read a textbook chapter or play the game compared to not participating in these learning activities. Conversely, for students with good prior knowledge, these activities might not yield additional improvements.

TABLE 4: Average MC and essay question scores in treatment and control groups

	MC questions		Essay questions	
Prior knowledge	poor	good	poor	good
Textbook	4.48	5.10	55.7	62.7
Game	4.25	3.56	55.9	65.0
Micro 1 students, prepared	3.60	5.13	44.8	58.5
Micro 1 students, unprepared	3.71	5.11	32.0	58.0
Non-Micro 1 students	4.05	7.00	32.6	62.7

Table 4 shows average multiple-choice and essay question scores for our textbook and game-play conditions, as well as for the three control groups, separately for those who judged their prior knowledge as poor and good, respectively. For students with little prior knowledge, we observe that both multiple-choice and essay scores in the textbook and game conditions are higher than in the three control conditions. When statistically comparing the treatment groups with the control groups (using non-parametric Wilcoxon ranksum tests), the positive learning effects for students with poor prior knowledge are highly significant only for essay questions ($p < 0.001$ and $p < 0.001$ for textbook vs. controls and game vs. controls, respectively), but weakly or not significant for multiple-choice scores ($p = 0.059$ and $p = 0.338$ for textbook vs. controls and game vs. controls, respectively). We interpret the positive effects as evidence for the effectiveness of both textbook and game play in facilitating learning. No learning effects were observed for the group of students with good prior knowledge, as is confirmed by a series of Wilcoxon ranksum tests comparing these students in the two treatment and control conditions ($p > 0.3$ in all tests with one exception).[2]

5.4 Game behavior and learning outcomes

In order to capture the behavior of students within the game, we obtained the following data from the game play recordings:

- *LevelAttempts*: The number of attempts needed to achieve the primary objectives of a game level. If a student does not achieve the primary objectives in a level play, he has to repeat the level until he/she does.
- *LevelPlaythroughs*: How often a level was played through. This measures how often a student went back and replayed a level he/she has already accomplished.
- *VideosPlayed*: How often instruction videos were played (for more than 5 seconds) in the game session.
- *TotalTimeVideos*: The total time (in minutes) spent in the game's media library, not including the compulsory initial instruction video.
- *TotalTimeGraphs*: The total time (in minutes) spent looking at explanatory graphs.

In addition to these parameters, which ought to represent the deliberate engagement of the student with the game, we also obtained measures for pure activity (*NumClick*, the average number of mouse clicks during the first playthrough of the levels, and *NumHover*, the number of times a participant interacted with one of the instructional graphs) and accuracy of play (*AvgPrimObj* and *AvgSecObj*, the average percentage of primary and secondary objectives completed over all levels, respectively), which we will use as additional controls.

[2] A statistical comparison between the multiple-choice question performance of students with good prior knowledge in the Game treatment and the baseline conditions indicates a *negative* effect (3.56 vs. 5.1, $p = 0.001$). In principle, such an effect might have many reasons, for example it may root in a negative effect of deeper knowledge on performance in ambiguously worded multiple-choice questions. A separate study designed to address this question would be necessary to confirm whether this unexpected effect is a statistical fluke or not.

TABLE 5: Tobit regressions on the effect of game play on test results

Dependent	Correct MC questions			Total marks in essay questions		
Model	(1)	(2)	(3)	(4)	(5)	(6)
Constant	1.37	3	4.24	40.58**	109.28***	196.00***
	[1.26]	[2.23]	[4.08]	[17.72]	[29.56]	[50.69]
LevelAttempts	0.03	0.02	0.01	-0.21	-0.44	-0.53
	[0.04]	[0.04]	[0.04]	[0.50]	[0.47]	[0.47]
LevelPlaythroughs	0.18**	0.19***	0.20**	2.60**	2.97***	2.48**
	[0.07]	[0.07]	[0.08]	[0.98]	[0.91]	[0.94]
VideosPlayed	0.04	0.05	0.04	1.84*	2.16**	1.3
	[0.07]	[0.07]	[0.08]	[0.96]	[0.96]	[0.97]
TotalTimeVideos	-0.07	-0.05	-0.05	-2.09**	-1.89**	-2.72***
	[0.06]	[0.06]	[0.07]	[0.80]	[0.80]	[0.85]
TotalTimeGraphs	0.04	0.09	0.07	0.73	-0.17	-1.7
	[0.14]	[0.15]	[0.16]	[1.99]	[2.03]	[2.03]
NumClick		-0.01	-0.01		-0.16**	-0.12
		[0.01]	[0.01]		[0.07]	[0.07]
TimeHover		-0.01	0		0.43	-0.49*
		[0.02]	[0.02]		[0.28]	[0.25]
AvgPrimObj		-0.01	-0.01		-0.3	-1
		[0.02]	[0.02]		[0.24]	[1.92]
AvgSecObj		0.06	0.06		-1.39	0.37
		[0.14]	[0.15]		[1.87]	[0.27]
Gender			-0.28			-1.55
			[0.54]			[6.69]
Age			-0.05			-2.67
			[0.14]			[1.74]
International			-0.1			-3.7
			[0.47]			[5.09]
Commerce/Economics			-0.03			-13.56**
			[0.50]			[6.23]
N	42	42	42	42	42	42
Left-(non-)right-censored	0(41)1	0(41)1	0(41)1	0(42)0	0(42)0	0(42)0
LogLL	-70.8	-69.3	-68.9	-182.4	-178.4	-175.3

Note: Standard errors are given in brackets. *,**, and *** denote significance at the 10%, 5%, and 1% level, respectively.

Based on data from all the participants in our game conditions, Table 5 reports the results from regressions of the number of multiple-choice questions solved and marks obtained in the essay questions. Models (1) and (4) regress our two dependents on the main game behavior parameters discussed above. In Models (2) and (5) we add the activity and accuracy controls, and Models (3) and (6) check the robustness of the estimates with respect to the same demographic characteristics used in Table 3 analysis. For test results on multiple-choice questions, how often a participant voluntarily replayed the game levels is quite important. While causality is not straightforward here, this could indicate that game play elements are most important for repetitive learning. For the essay questions on the other hand, both watching instructional videos and engaging in game play seem to matter. This indicates that to score high on deeper learning questions, using direct instructions might be important too. These results are very robust across different specifications, both in terms of size and significance of the effects. We also find two spurious effects for the activity and accuracy independents, which are, however, not robust across specifications.

In sum, we find support for Hypothesis 5 in that game behavior and test performance are related, and differently so for different types of learning and learning assessment. A question by question analysis allows us to dig deeper into these relationships. Using Bloom's taxonomy (Bloom, Engelhart, Furst, Hill and Krathwohl, 1956), we first ranked the five essay questions based on the six learning objectives they aim to achieve, i.e., knowledge, comprehension, application, analysis, synthesis, evaluation. Questions 1 and 2 touched on the first two and three objectives, respectively; Questions 3 and 4 focused on the first four objectives, while Question 5 addressed all six of them. With this classification in mind, we re-run the specifications from Table 5 separately for each essay question, and each time we included only the variables corresponding to the pertinent video-game level (regression results not included here). As expected, we find that the number of voluntary level replays has a crucial role at low and medium levels of taxonomy (i.e., first three essay questions), while for low taxonomy (Questions 1-2) the accuracy of play (i.e., $AvgSecObj$) and videos also seem to matter. For high levels of taxonomy on the other hand (i.e., Question 5 which touches on all six taxonomy levels), graphs are very important, while the number of level attempts appears negatively associated with the test scores.

5.5 *Students evaluation of the learning experience*

After the experiment (i.e., after the test had been completed, but before any results were announced and payoffs were handed out), all participants were asked to fill in a post-experimental questionnaire. Attitudes were elicited on either yes/no questions or 7-level Likert-type scales with answer options ranging from "Strongly disagree" (coded 1) to "Strongly agree" (coded 7). Answers were not compulsory, but most participants answered all questions. The upper part of Table 6 reports the answers given to the questions targeting students' perceptions of the learning experience. Since Likert scales elicit ordinal data (in the sense that the difference between two consecutive answer levels might not be the same for all levels), we only refer to whether an item was ranked higher or lower in one treatment condition than in the other. We find that students evaluated the textbook/game sessions similarly in terms of understanding what was required of them, usefulness of the textbook/game in absorbing the topic, and how easy it was for them to read the textbook or play the game (Wilcoxon ranksum tests comparing textbook and game answers yield p-values of 0.507, 0.127, and 0.828, respectively). However, participants enjoyed playing the game significantly more than reading the textbook chapter (Wilcoxon p-value = 0.002). In terms of "spreading the word", participants in both conditions are about equally likely to recommend the textbook/game to a fellow student in Microeconomics I (81% vs. 91%, respectively; χ^2-test p-value = 0.409), but were (weakly) significantly more likely to recommend the game (rather than the textbook) to a friend (50% and 77%, respectively; χ^2-test p-value = 0.088).

In Section 3 we hypothesized that a possible reason for gender effects in terms of the relative performance after game-play versus textbook could be that males enjoy video-games more than females. In Table 6 we report enjoyment ratings separately for women and men. We indeed detect a small gender effect. While both men and women show an increase in enjoyment when playing the game rather than reading the textbook (Wilcoxon ranksum test p-values of 0.052 and 0.004, respectively), men like the game more than women (p=0.056), while there is no statistically significant difference between the two groups in the textbook

TABLE 6: Results from the post-experimental questionnaire on learning experience perceptions

	Textbook		Game	
	Mean	(Std. Dev.)	Mean	(Std. Dev.)
Easy to understand requirements (1-7)	5.72	(1.12)	5.57	(1.11)
Helped improve understanding of topic (1-7)	5.97	(0.87)	5.64	(1.08)
Easy to read / play (1-7)	5.85	(0.96)	5.74	(1.19)
Enjoyed reading / playing (1-7)	4.72	(1.61)	5.76	(1.10)
Females	4.00	(2.00)	5.44	(1.03)
Males	4.93	(1.44)	5.96	(1.11)
Suggest to Micro 1 student (y/n)	yes: 85%, no: 15%		yes: 91%, no: 9%	
Suggest to a friend (y/n)	yes: 56%, no: 44%		yes: 77%, no: 23%	
Usefulness of elements	Score (1-7)		Rank (1-5)	
Textbook	Mean	(Std. Dev.)	Mean	(Std. Dev.)
Main sections	5.92	(0.74)	2.54	(1.31)
Exercises	5.56	(1.21)	3.74	(1.14)
Thinking as an Economist sections	5.56	(1.23)	3.67	(1.28)
Examples	6.08	(0.87)	1.67	(0.81)
Graphs	5.74	(0.99)	3.36	(1.31)
Game				
Videos	5.86	(0.93)	1.66	(1.09)
Score system	5.07	(1.18)	4.13	(1.18)
Personalised Feedback	5.60	(1.33)	3.55	(1.04)
Gameplay elements	5.76	(1.05)	3.38	(1.17)
Interactive graphs	6.26	(0.80)	2.28	(1.06)

Note: The exact statements/questions were: "It was easy to understand what was required of me", "The material helped improve my understanding of the topic.", "I found it easy to read and understand the textbook/ work with the interactive elements of the game.", "I enjoyed reading the textbook / playing the game.", "Would you suggest this textbook / game to a colleague studying ECON1101?", "In general, would you suggest this textbook / game to a friend?", and "I found ... useful in my understanding of the topic.".

treatment (p=0.1954).

In order to explore the correlation between the perception of the learning environment and performance in the tests, we regressed the test performance on the six perception indicators (regression results not included here). We did not find any evidence that those who perceive the environment more positively would achieve better scores, neither in general nor for a specific gender. Notwithstanding any possible selection effects underlying these results (those who like an environment – like the game – might not be the better students), this finding does not lend evidence to our Hypothesis 6, but is also particularly interesting with respect to the gender effect: While women enjoy the game less than men, they seem not to have to actually enjoy the educational video-game as much as men do in order to learn within this environment. This finding is consistent with the Hypothesis 4 results, where we do not find gender effects in terms of relative exam performance in the game treatment.

A further set of questions asked participants to rate and rank the different educational elements of the textbook/game in terms of their usefulness. Asking for ratings gives a relatively independent evaluation, while the ranking forces comparisons between elements and yields statistically more sensitive data (since students had to decide on a different rank for each element). The answers, reported in the lower part of Table 6, allow us to get a more detailed view of how students perceive the two learning experiences.

For the textbook, we obtain exactly the same order of average preferences from the ratings and from the rankings. Examples are evaluated as most useful, the Main Sections came in second, while the other three elements of Graphs, Exercises, and "Thinking as an Economist" sections are rated and ranked about equally. Statistically, this order is confirmed by a battery of Wilcoxon Matched Pairs Signed Ranks tests on both scores and ranks. For ratings, only Examples receives significantly higher scores than Exercises and Thinking sections (both p-values smaller than 0.05), and Main Sections receive weakly significantly higher scores than Exercises (p=0.099), while all other comparisons yield non-significant differences (all p-values larger than 0.12). The more sensitive ranking data yields a clear order of Examples being ranked significantly higher than Main Sections, and both ranked significantly higher than the other three element types (all p-values smaller than 0.02), while there are no statistical differences among the ranks assigned to other elements (all p-values larger than 0.23).

Interestingly, the answers for the game environment look different. Statistically, the rating data yields a clear ordering of Interactive Graphs $>$ Videos $=$ Game-play elements $=$ Personalized Feedback $>$ Score System (Wilcoxon Matched Pairs Signed Ranks test p-values smaller than 0.02 for all different average ratings and p-values of more than 0.29 for all comparisons where we don't find differences). The ranking data, on the other hand, yield an ordering of Videos > Interactive Graphs > Game-play elements = Personalized Feedback > Score System (p-values smaller than 0.05 for all implied differences, p-values of 0.618 for the single implied indifference). Thus, with ratings, Interactive Graphs are at first place and Videos come in second, while in the (forced) rankings Videos are considered as more useful than the Interactive Graphs. Game play and Personalized Feedback follow, statistically indistinctive, on 3rd and 4th place, while the Score System is ranked least useful.

All in all, when learning from a textbook, the most useful elements seem to be the examples (that is, the application of the material to real problems), while the actual explanations of the topic are ranked as less useful. In the video-game environment, which is specifically designed around examples that can be played out, the explanations of the topics (videos and graphs) are ranked most useful. One explanation might be that both environments set different reference points. In textbooks, explanations are the standard and insightful examples the exception. But when approaching a game, the user expects all the game play elements, and is potentially positively surprised when encountering useful explanatory learning tools.

6 Conclusions

This paper presents experimental evidence on the effects of playing an educational video-game *vs.* the traditional method of textbook learning on both cognitive and affective (behavioral) outcomes. To this effect, we use a game designed to support teaching introductory microeconomics at the undergraduate level. We find no evidence that reading a textbook dominates playing the video-game in terms of exam performance, neither for multiple-choice nor for essay questions. Moreover, there seems to be no gender bias and no incentive effect for final scores. Different than in previous studies, we find that game behavior is correlated with exam performance, and that this relation differs with the types of learning and assessment. Students enjoyed the video-game considerably more than reading the textbook, but did not perceive any differences between the two learning tools in terms of understanding requirements or usefulness. Finally, enjoyment of the game is particularly strong for males, but does not lead to a differential effect between genders in terms of test performance.

In sum, based on a controlled laboratory experiment, the results from our study suggest that a video-game might be a cost-efficient, high-quality alternative or complementary teaching method to traditional textbook and classroom teaching. However, our experiment only presents a first case study, based on a limited number of observations, and testing a specific video-game on a specific topic in the area of undergraduate economics. More experiments, including proper design and control groups, are needed to explore the effectiveness of video-games as a teaching and learning tool in general, in other areas. Future research will aim to enrich the body of evidence, and to test the robustness of our results. Our paper is a case in point that shows how such studies can be designed in order to provide clean data, prevent selection, and allow conclusions about causality.

References

Anderson, P. H. and Lawton, L. (2009), 'Business simulations and cognitive learning: Developments, desires, and future directions', *Simulation and Gaming* **40(2)**, 193–216.

Anderson, P. and Lawton, L. (1992), 'The relationship between financial performance and other measures of learning on a simulation exercise', *Simulation and Gaming* **23**, 326–340.

Bangert-Drowns, R., Kulik, C., Kulik, J. A. and Morgan, M. (1991), 'The instructional effect of feedback in test-like events', *Review of Educational Research* **61(2)**, 213–238.

Barab, S. A., Gresalfi, M. and Ingram-Goble, A. (2010), 'Transformational play: Using games to position person, content, and context', *Educational Researcher* **39(7)**, 525–536.

Bell, B. and Kozlowski, S. (2008), 'Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability', *Journal of Applied Psychology* **93**, 296–316.

Bennett, E., Mims, G. and McKenzie, B. (2002), Assessing distributed learning: Student perceptions and future directions, *in* 'Society for Information Technology and Teacher Education International Conference', Vol. 1, pp. 2379–2382.

Bettinger, E. and Long, B. T. (2007), Institutional responses to reduce inequalities in college outcomes: Remedial and developmental courses in higher education, *in* S. Dickert-Conlin and R. Rubenstein, eds, 'Economic Inequality in Higher Education: Access, Persistence, and Success', Russell Sage Foundation Press, New York.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. and Krathwohl, D. R. (1956), *Taxonomy of educational objectives: The classification of educational goals, Handbook I: Cognitive Domain*, Longmans and Green, New York.

Brown, K. G. and Ford, J. K. (2002), Using computer technology in training: Building an infrastructure for active learning, *in* K. Kraiger, ed., 'Creating, Implementing, and Maintaining Effective Training and Development: State-of-the-art Lessons for Practice', Jossey-Bass, San Francisco, pp. 192–233.

Burns, A. C., Gentry, J. W. and Wolfe, J. (1990), A cornucopia of considerations in evaluating the effectiveness of experiential pedagogies, *in* J. W. Gentry, ed., 'Guide to business gaming and experiential learning', Nichols/GP Publishing, East Brunswick, NJ, pp. 253–278.

Chirieac, D., Burns, M. and Case, T. (2000), 'The impact of gender and time on attitude toward to the internet and computing technology', Georgia Southern University Department of Information Systems and College of Business Administration, Statesboro.

Croson, R. and Gneezy, U. (2009), 'Gender differences in preferences', *Journal of Economic Literature* **47(2)**, 448–474.

Csikszentmihalyi, M. (1990), *Flow: The Psychology of Optimal Experiences*, New York: Harper and Row.

Dickie, M. (2006), 'Do classroom experiments increase learning in introductory microeconomics?', *Journal of Economic Education* **37(3)**, 267–288.

Donovan, C., Figlio, D. and Rush, M. (2006), 'Cramming: The effects of school accountability on college-bound students', NBER Working Paper, No. 12628.

Emerson, T. L. N. and Taylor, B. A. (2004), 'Comparing student achievement across experimental and lecture-oriented sections of a principles of microeconomics course', *Southern Economic Journal* **70(3)**, 672–693.

Emerson, T. L. N. and Taylor, B. A. (2007), 'Interactions between personality type and the experimental methods', *Journal of Economic Education* **38(1)**, 18–35.

Emerson, T. L. N. and Taylor, B. A. (2010), 'Do classroom experiments affect the number of economics enrollments and majors? a study of students in the united states', *International Review of Economic Education* **9(2)**, 43–58.

Faria, A. J. (2001), 'The changing nature of business simulation/gaming research', *Simulation and Gaming* **32**, 97–110.

Faria, A. J. and Wellington, W. J. (2004), 'A survey of simulation game users, former-users, and never-users', *Simulation and Gaming* **35**, 178–207.

Figlio, D. N., Rush, M. and Yu, L. (2010), 'Is it live or is it internet? experimental estimates of the effects of online instruction on student learning', NBER Working Paper, No. 16089.

Fischbacher, U. (2007), 'z-Tree: Zurich Toolbox for Ready-made Economic experiments', *Experimental Economics* **10(2)**, 171–178.

Frank, R., Jennings, S. and Bernanke, B. (2012), *Principles of Microeconomics, 3rd Edition*, McGraw Hill, Australia.

Frese, M., Brodbeck, F. C., Heinbokel, T., Mooser, C., Schleiffenbaum, E. and Thiemann, P. (1991), 'Errors in training computer skills: On the positive function of errors', *Human Computer Interaction* **6**, 77–93.

Fripp, J. (1993), *Learning Through Simulation*, McGraw-Hill, London.

Garris, R., Ahlers, R. and Driskell, J. E. (1981), 'Games, motivation, and learning: A research and practice model', *Simulation and Gaming* **33**, 441–467.

Gneezy, U., Niederle, M. and Rustichini, A. (2003), 'Performance in competitive environments: Gender differences', *Quarterly Journal of Economics* **118(3)**, 1049–1074.

Gosen, J. J. and Washbush, J. (2004), 'A review of scholarship on assessing experiential learning effectiveness', *Simulation and Gaming* **35**, 270–293.

Gregory, M. Y. (1997), Gender differences: An examination of computer mediated communication, *in* 'The Annual Meeting of the Southern States Communication Association, ERIC No: ED410604'.

Greiner, B. (2004), An Online Recruitment System for Economic Experiments, *in* K. Kremer and V. Macho, eds, 'Forschung und wissenschaftliches Rechnen 2003, GWDG Bericht 63', Ges. für Wiss. Datenverarbeitung, Göttingen, Germany, pp. 79–93.

Hoska, D. (1993), Motivating learners through cbi feedback: Developing a positive learner perspective, *in* J. V. Dempsey and G. Sales, eds, 'Interactive instruction and feedback', Educational Technology, Englewood Cliffs, NJ, pp. 105–132.

Jackson, L. A. and Ervin, K. (2001), 'Gender and internet: Women communicating and men searching', *Sex Roles: Journal of Research* **44(6)**.

Jazwinski, C. (2001), *Gender Identities on the World Wide Web: Learning and Teaching on the www*, Academic Press, San Diego.

Jonassen, D. (2002), Integration of problem solving into instructional design, *in* R. Reiser and J. Dempsey, eds, 'Trends and Issues in Instructional Design and Technology', Merrill/Prentice Hall, Upper Saddle River, NJ, pp. 107–120.

Kearsley, G. (2000), *Online Education: Learning and Teaching in Cyberspace*, Wadsworth Publishers, USA.

Lean, J., Moizer, J., Towler, M. and Abbey, C. (2006), 'Simulations and games: Use and barriers in higher education', *Active Learning in Higher Education* **7**, 227.

Lucas, K. and Sherry, J. L. (2004), 'Sex differences in video game play: A communication-based explanation', *Communication Research* **31(5)**, 499–523.

Malone, T. W. (1981), 'Toward a theory of intrinsically motivating instruction', *Cognitive Science* **4**, 333–369.

Niederle, M. and Vesterlund, L. (2007), 'Do women shy away from competition? do men compete too much?', *Quarterly Journal of Economics* **122(3)**, 1067–1101.

Niederle, M. and Vesterlund, L. (2011), 'Gender and competition', *Annual Review of Economics* **3(1)**, 601–630.

Saunders, P. M. (1997), 'Experiential learning, cases, and simulations in business communication', *Business Communication Quarterly* **50(1)**, 97–114.

Schreiber, D. and Berge, Z. (1998), *Distance Training: How Innovative Organizations are Using Technology to Maximize Learning and Meet Business Objectives*, JosseyBass Press, San Francisco.

Shute, V. J. (2008), 'Focus on formative feedback', *Review of Educational Research* **78(1)**, 153–189.

Sitzmann, T. (2011), 'A meta-analystic examination of the instructional effectiveness of computer-based simulation games', *Personnel Psychology* **64**, 489–528.

Sitzmann, T., Brown, K. G., Casper, W., Ely, K. and Zimmerman, R. D. (2008), 'A review and meta-analysis of the nomological network of trainee reactions', *Journal of Applied Psychology* **93**, 280–295.

Sloman, J. (2002), International trade game, Classroom game, University of the West of England, Available at http://www.economicsnetwork.ac.uk/showcase/sloman_game.

Squire, K. (2006), 'From content to context: Videogames as designed experience', *Educational Researcher* **35(8)**, 19–29.

Sugrue, B. (1995), 'A theory-based framework for assessing domain-specific problem solving ability', *Educational Measurement: Issues and Practice* **14(3)**, 29–36.

Tennyson, R. D. and Jorczak, R. L. (2008), A conceptual framework for the empirical study of instructional games, *in* H. F. O'Neil and R. S.Perez, eds, 'Computer Games and Team and Individual Learning', Elsevier, Oxford, UK, pp. 39–54.

Venkatesh, V. and Morris, M. G. (2000), 'Why don't men ever stop to ask for directions? gender, social influence, and their role in technology acceptance and usage behaviour', *Management Information Systems Quarterly* **24(1)**, 115—139.

Vorderer, P., Hartmann, T. and Klimmt, C. (2006), Explaining the enjoyment of playing video games: The role of competition, *in* D. Marinelli, ed., 'ICEC Conference Proceedings 2003: Essays on the Future of Interactive Entertainment', Carnegie Mellon University Press, Pittsburgh, PA, pp. 107–120.

Washbush, J. and Gosen, J. (2001), 'An exploration of game-derived learning in total enterprise simulations', *Simulation and Gaming* **32**, 281–296.

Wideman, H. H., Owston, R. D., Brown, C., Kushniruk, A., Ho, F. and Pitts, K. C. (2007), 'Unpacking the potential of educational gaming: A new tool for gaming research', *Simulation and Gaming* **38**, 10–30.

Wolfe, J. (1990), The evaluation of computer-based business games: Methodology, findings, and future needs, *in* J. W. Gentry, ed., 'Guide to Business Gaming and Experiential Learning', Nichols/GP Publishing, East Brunswick, NJ, pp. 279–300.

Wright, J. C., Huston, A. C., Vandewater, E. A., Bickham, D. S., Scantlin, R. M., Kotler, J. A., Caplovitz, A. G. and Lee, J. (2001), 'American children's use of electronic media in 1997: A national survey', *Applied Developmental Psychology* **22(1)**, 31–47.