CITATION:

Brom, C., Levčík, D., Buchtová, M., Klement, D. (2015) Playing Educational Micro-games at High Schools: Individually or Collectively? Computers in Human Behavior, Vol. 48, July (2015), pp. 682-694, DOI information: 10.1016/j.chb.2015.02.025

Title page

1) Full title:

Playing Educational Micro-games at High Schools: Individually or Collectively?

2) Authors:

Cyril Brom

Faculty of Mathematics and Physics, Charles University in Prague

Malostranské nám. 25, 118 00, Prague, the Czech Republic

brom@ksvi.mff.cuni.cz

David Levčík

Institute of Physiology, Academy of Sciences of the Czech Republic v.v.i.

Vídeňská 1083, 14220 Prague 4, Czech Republic

david.levcik@fgu.cas.cz

Michaela Buchtová

Faculty of Mathematics and Physics, Charles University in Prague

Malostranské nám. 25, 118 00, Prague, the Czech Republic

Faculty of Arts, Charles University in Prague

U kříže 8, 15800 Prague 5, Czech Republic

michaela.buchtova@ff.cuni.cz

Daniel Klement

Institute of Physiology Academy of Sciences of the Czech Republic v.v.i.

Vídeňská 1083, 14220 Prague 4, Czech Republic

dklement@biomed.cas.cz

3) Corresponding author:

Cyril Brom

Faculty of Mathematics and Physics, Charles University in Prague, Room 312,
Malostranske Namesti 25, Prague, 11800, Czech Republic.

E-mail: brom@ksvi.mff.cuni.cz

Tel: (420) 221 914 216; Fax: (420) 221 914 281

4) Source of funding

Cyril Brom was partially supported by project nr. P407/12/P152 supported by Czech Grant Science Foundation (GA ČR). David Levčík was partially supported by IGA MZ CR grant nr. NT13386. Michaela Buchtová was partially supported by project Digital Technologies in Education from Grant Agency of Charles University (GAUK), project no. 581012. The OPB game was developed as part of the project CZ.2.17/3.1.00/ 31162 financed by the European Social Fund and the Municipality of Prague.

5) Acknowledgement

We would like to thank all students who participated in the experiment and also to the teachers and school administrators who made this experiment possible. Special thanks go to Michal Preuss and František Šusta, who helped with development of the OPB game, and to Tereza Nekovářová, who helped with the experiment design. We also thank Uri Wilensky for allowing us to use Netlogo in this research.

The human data were collected with APA ethical principles in mind.

Playing Educational Micro-games at High Schools: Individually or Collectively?

Abstract

The effectivity of learning by playing serious games is increasingly subject to research, but information about how these games should actually be used in classes is limited. In this explorative study with between-subject design (*N* = 166; high school students), we investigated the effectivity of playing two different micro-games in two different ways. After an expository lecture, either students played a game individually at computers ("individual play"), or the teacher played it, while showing it to the class on a projector and prompting the students on how to proceed with the game ("collective play"). Results indicated that the two modes of play were nearly comparable as concerns immediate and one month delayed learning gains, as well as subjective evaluation of educational experience. There were only two notable differences. First, immediate test scores for factual questions, but not transfer questions, for one of the games were higher for the individual play (medium effect size). Second, this difference was accompanied by a higher enjoyment in the better performing group (small to medium effect size). The results support the idea that collective play, which is easier to implement in schools, is a method that should be considered in educational design and future research.

Keywords: educational games, collective play, learning effects, secondary education, classrooms

Article

1 Introduction

Some schools have gradually adopted educational digital games for use (Wastiau, Kearney, & den Berge, 2009; Williamson, 2009; Huizenga, Admiraal, & Ten Dam, 2013; Proctor & Marks, 2013) but these and also other educational institutions reported multiple issues (De Grove, Bourgonjon, & Van Looy, 2012; Kenny & McDaniel, 2011; Ketelhut & Schifter, 2011; Sisler & Brom, 2008; Wastiau et al., 2009, Ch. 7). Therefore, the games' acceptance is not always guaranteed and the issue of how to integrate them effectively into formal education remains an important question.

In a well-equipped school, with a committed teacher, students, parents, and school board, it is often possible to play games for educational purposes even over prolonged periods of time (Watson, Mong, & Harris, 2011; see also Gjedde, 2013; Wastiau et al., 2009). However, in many cases, in a moderately-equipped school, with an average teacher somewhat interested in game-based learning (but with other stakeholders who are largely uninformed), certain problems can emerge. Besides scepticism on the part of some of these stakeholders (Bourgonjon, Valcke, Soetaert, De Wever, & Schellens, 2011; Bourgonjon, Valcke, Soetaert, & Schellens, 2010; De Grove et al., 2012), there are potential problems with integrating games into curricula, issues with long learning curves (for both teachers and students), in situ technical problems, issues with implementing sustainable support for teachers, and cost limitations (e.g., Bourgonjon et al., 2010; De Grove et al., 2012; Egenfeldt-Nielsen, 2005; Ketelhut & Schifter, 2011; Klopfer, 2008; Wastiau et al., 2009; Williamson, 2009). For instance, it may be complicated for learners who do not play games regularly to learn how to control a complex game (Bourgonjon et al., 2010; Egenfeldt-Nielsen, 2005).

Typically, games also have to be played in a computer laboratory. In many schools, it may not be possible for teachers to take their class to the lab whenever they want because the lab is a shared resource (Kebritchi, 2010, p. 261; Watson & al., 2011; p. 473; see also Klopfer, 2008; Ch. 6). Coping with fixed lesson lengths (usually 45 – 60 minutes) is also a problem for many teachers; as is accommodating long game-play sessions within an overloaded curriculum (e.g., De Grove et al., 2012; Sandfrod, Uliscak, Facer, & Rudd, 2006).

However, many of these issues can be addressed. For example, there are attempts to overcome problems with access to a computer laboratory by using mobile devices specifically dedicated to learning through game playing (Klopfer, 2008) or by using an interpersonal computer (Szewkis et al., 2011).

One approach to overcoming both the technological and curricular-congruency problems is the use of *micro-games* (e.g., Brom, Preuss, & Klement, 2011). Use of single-player micro-games in the context of secondary education is in the scope of this paper.

Micro-games are “relatively simple computer games that do not require special skills to play and that challenge players with clearly-defined goals reachable within minutes or tens of minutes of game-play” (Brom et al., 2011, p. 1979). Unlike commercial-off-the-shelf games, they do not require up-to-date hardware, they fit well within short school lessons, and they are often directly created to cover a curricular topic. Because of their simplicity, developers can construct user interfaces that can be mastered within a matter of minutes. The drawback is that these games may be cognitively more superficial compared to games played over prolonged periods and/or repeatedly. Therefore, they may be less educationally effective (cf. Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013, p. 259). Nevertheless, micro-games can be, similarly to educational simulations, particularly useful as supplements to traditional lessons.

A single-player micro-game still, however, has to be played in a computer laboratory, or every student has to have a mobile device equipped with the game. Both approaches require expensive resources shared by many teachers. “Minor” technical issues, such as forgotten passwords or empty batteries, can also arise, which is troubling because solving these technical issues can take substantial time away from the class period (cf. Klopfer, 2008).

Alternatively, the teacher can play the micro-game on a computer directly in the classroom, via a projector, and prompt students in the class with questions on how to proceed in the game. Collectively, they can decide on the next steps (i.e., the students do not play the game individually: each with one device; instead, there is just one device controlled by the teacher). Such an approach is technically easier and cheaper to implement in real schools compared to lab visits or mobile devices, because it requires just one working computer and projector. Based on our eight years of experience with game-based learning, this is the approach many teachers tend to adopt in the case of micro-games. However, intuition suggests that under such “collective” play, learners tend to be cognitively more passive compared to “individual” play, and cognitive passivity does not promote meaningful learning (e.g., Mayer, 2004).[1]

It is thus useful to known if such “collective” play is comparable to “individual” play in terms of cognitive and affective outcomes; or to what extent is it worse. The answer can

[1] Note that collective playing *per se* is not collaborative.

have practical consequences for usage of micro-games in schools. Many media comparison studies, in which one group received game-based instruction and the second group a comparable "traditional" lesson, have been conducted in the past (meta-analysed in Vogel et al., 2006; Sitzmann, 2011; Wouters & al., 2013[2]). The treatments used for comparison were diverse (All, Nunez Castellar, & Van Looy, 2014; Wouters & al., 2013) but we are unaware of any study that would use, in the control group, the same single-player game where played collectively by the whole class.

In this paper, we present an exploratory study that investigates whether learning gains and subjective evaluation of the micro-game played differs if each student plays the micro-game individually in a lab, as opposed to collectively with a teacher as part of a small class (i.e., up to 15 students). The game is used as a supplement after a traditional expository lecture. To attempt at partial generalization, the study uses two different micro-games (the topics: genetics, animal learning). The study was conducted in schools as part of regular education programmes. We recruited 9 high school classes (10th or 11th grade) in two different, above-average, urban high schools in the Czech Republic and used stratified randomization (with the stratum being a class). Immediate and one-month delayed knowledge tests were administered and enjoyment and self-estimated learning assessed. Quantitative instruments were supplemented with informal in-class observations.

[2] The meta-analyses demonstrated the modest superiority of educational games but with some caveats, such as evidence of publication bias (Sitzmann, 2011) or diminishing positive impact of games in studies with randomization (Wouters et al., 2013).

2 Study Background

Educational computer games are multimedia learning materials. In this study, we adopt a view of multimedia learning as *active knowledge construction* (Mayer, 2009; pp. 17-19), as concerns declarative knowledge acquisition. In this view, learners construct a coherent mental model of a subject based on the learning material presented and their prior knowledge. Model construction is a personal process and learners must engage in it *actively* to create their own *individual* knowledge. Teachers serve as “cognitive guides” assisting the learners during the knowledge construction.

In this view, cognitive activity rather than behavioural/physical activity is important for effective (declarative) knowledge acquisition (Mayer, 2004; Mayer, 2009; p. 23). Behavioural activity may nevertheless serve as a trigger for instigating cognitive activity. This is important for educational games, which are highly interactive and interactivity is one of the key features for promoting behavioural activity; but not necessarily cognitive activity.

This view is adopted by the prominent theory of multimedia learning, Cognitive Theory of Multimedia Learning (CTML; Mayer, 2009). Capitalizing on Baddeley’s classical memory model (Baddeley, Eysenck, & Anderson, 2009) and Dual Coding Theory (Clark & Pavio, 1991), CTML posits that multimedia information is processed by learners through two separate cognitive channels (verbal and visual), organized in their working memory into coherent models and integrated with prior knowledge “stored” in their long-term memory. This process’ efficiency depends on the level of a learner’s active cognitive participation (Mayer, 2009; Moreno, 2010; Moreno & Mayer, 2007), on available cognitive capacity (Sweller, 1999) and various other learner- and environment-related contextual factors (Moreno, 2005; Moreno & Mayer, 2007).

The active-learning view is also reflected in the Integrated Model of Multimedia Interactivity (INTERACT) (Domagk, Schwartz, & Plass, 2010), proposed by its authors as an explanatory framework for *interactive* multimedia learning studies. Two primary components of this six-component, process-oriented model are Behavioural Activities and Cognitive/Metacognitive Activities. Other components include Learning Environment, Learner Variables (trait-like), Emotion and Motivation (state-like), and Mental Model. Interactivity is represented by feedback loops among the majority of these components. This model explicitly informs us that behavioural activity alone is insufficient for deep-level cognitive processing, which is vital for mental model construction.

In this study, we are interested in using an interactive micro-game in two different ways after an expository lecture. The lecture is the same for both conditions and the game is used for the same purpose in both conditions: *reinforcing* and *integrating* (in the terms of Thomas & Hooper, 1991) part of the knowledge learnt in the lecture. However, there are between-condition differences as concerns the game's usage method.

In the "individual" condition, each student plays the game at one computer in a lab and the teacher serves as a coach for the whole class. From the perspective of the CTML and the INTERACT, the advantage of this approach is that students must frequently exert behavioural activity in order to proceed further in the game (because of the game's interactivity). Even though behavioural activity may not always trigger cognitive activity (some students may click their way through the game or devote their time to non-gaming activities; cf. Baker, Corbett, Koedinger, & Wagner, 2004), it is still probable that more students will be cognitively active in this condition compared to the "collective" play. A disadvantage of the individual play is that it is asynchronous: students proceed at their own pace and if the teacher needs to address the whole class, he or she will inevitably interrupt some of the learners, possibly causing a distraction and thus reducing cognitive activity.

In the “collective” condition, students sit in a regular classroom and the teacher stands in front of the whole class. The teacher initiates a game action on his/her computer, while showing it via a projector screen. Alternatively, the teacher may project on the screen a pre-prepared slide with a screenshot of the game. Visually, what students see would be very similar to what they would see in the “individual” condition; however, now actions have not been initiated by their individual behavioural activity. Unlike in the “collective” condition, the teacher can also address the whole class after every game action to comment on or contextualize what has happened on the screen and ask students how to proceed with the game. Apart from the disadvantage of this condition mentioned above (i.e., more students will likely be on off-task behaviour), this condition has also two advantages. First, those few who will be engaged by the teacher’s inputs may start an on-topic discussion and be consequently more cognitively active compared to a typical player in the “individual” condition. Second, the method is synchronous; with possibly lower distraction for students compared to the “individual” condition.

Therefore, the INTERACT and the CTML challenge the intuitive assumption that the “individual” condition is necessarily better for learning. These frameworks do not make clear predictions in favour of one of the conditions, but they nevertheless highlight that the “collective” play may be better than intuitively expected because it has two potential advantages: it is synchronous and it enables initiation of topical discussions. These properties promote cognitive activity, which may counterbalance the disadvantage of this condition caused by a higher behavioural passivity.

The INTERACT model and an extension of the CTML (Moreno, 2005; Moreno & Mayer, 2007) also inform us that cognitive activity actually used depends on various contextual factors, such as learners' affective state. In the multimedia learning field, it is currently a subject of discussion and research whether (and when) positive affect increases or decreases learning gains (e.g., Brom, et al., 2014; Plass, Heidig, Hayward, Homer, & Um, 2014): positive affect can arguably increase cognitive capacity actually used, thereby boosting learning, but also cause distraction, thereby impeding learning. Because the present study is an explorative one, pinning down the complex relationship between affect and learning in the context of our two conditions falls outside our present scope. Nevertheless, we also report initial summative data on students' enjoyment of learning in our two conditions, because these data can be important for teachers deciding on which approach to adopt.

3 Research Question

Should the teacher in a "typical" high school class use a micro-game in the "individual" or "collective" way? Because of the lack of previous research on this topic and because the theoretical frameworks do not clearly favour either of the conditions, this study is designed as an explorative one. We put forward no explicit hypothesis but rather a research question:

Is an educational method capitalizing on "individual" play (and employing a micro-game and other multimedia materials as delivery media) more, equally or less effective compared to the educational method capitalizing on "collective" play (with the same delivery media)?

Learners who perform well on the transfer test probably had the capacity for deep learning, which comes through active cognitive processing (cf. DeLeeuv & Mayer, 2008) (assuming low prior knowledge). Therefore, transfer tests, administered immediately after the intervention and also one month-later, will serve as our primary measuring instrument. They will be supplemented by retention tests, feedback questionnaires and informal observations of students' overall activity.

4 Method

4.1 Experiment Design

This explorative study compares (a) a teaching session consisting of an expository lecture (40 min.) followed by students' playing a single-player micro-game (30 min.) individually in a computer lab ("I" condition – individual) with (b) a teaching session consisting of the same expository lecture followed by playing the same game by the teacher, while showing it on a projector ("C" condition – collective). In schools, these modes can be implemented within two consecutive school hours. In the "C" condition, the teacher prompts students on how to proceed in the game, as detailed in Sec. 2. In both conditions, a micro-game is used as a supplement to the lecture. Under both conditions, during the game playing part, the teacher occasionally shows supplementary multimedia materials using a projector (such as screenshots from the game along with short expository texts). The teacher also provides short verbal inputs in order to explain what happens in the game and relate it to the information from the lecture. We assume that a class has up to 15 students, so that useful teacher – student interactions can occur.

The time allotment was the same for both conditions. The teacher was always the same: a member of the experimental team (he was the first author of this paper). The study recruited nine high school classes from two different high schools. We used two different micro-games (and therefore two different expository lectures and sets of multimedia materials) of comparable complexity. Five classes (three from one school and two from the other school) played one game and the remaining classes played the other game. Each class was randomly divided into the “C” and “I” condition after the expository lecture.

We administered background questionnaires, immediate post-tests and one-month delayed post-tests. To avoid cueing the students on what should be remembered, we did not administer pre-tests (see, e.g., Judd, Smith, & Kidder, 1991).[3] Supplementary qualitative data on students’ overall behavioural activity were produced by informal observations.

[3] To obtain baseline for the knowledge tests, we recruited 38 students from the same classes (absent during the experimental day) as naïve participants (see Sec. 4.5 for details) and administered them the tests. There were clear differences between the scores of naïve participants and the one-month delayed test scores of experimental participants (Cliff’s delta in the range from .343 to .609, i.e., medium to large effect sizes; *p* in the range from .008 to .0002). This indicates that those who underwent one of our treatments learnt by it and did not forget all the acquired information during the month after the experiment. We also have reasons to believe that the educational method combining an expository lecture with a micro-game is more beneficial than the lecture-only teaching method: in our previous study (Brom et al., 2011) with the same OPB game, where we compared the expository lecture followed by playing the game to a teaching session consisting of the same expository lecture followed by another short lecture that used only non-game multimedia learning materials, the game group outperformed the control group in one-month delayed knowledge tests (see also Sitzmann, 2011 for more on this point).

4.2 Participants

We recruited a total of 226 students (Mean age = 16.8 years, SD = 0.6 years, range 15-18 years; 10th or 11th grade) from nine different classes from two above-average urban high schools in the Czech Republic. Four classes were from one high school and five from another. One class had a specialization in ICT, two in natural sciences, two in humanities and four had no specialization. We excluded 60 students from the evaluation; those, who did not finish the delayed post-test. We included a total of 166 students in the analysis (88 males; 78 females). Participation in the experiment was part of the students' regular education; however, the students were not graded based on their performance on our knowledge tests/the game.

4.3 Micro-games and Multimedia Materials

We picked two curriculum-congruent topics, animal learning and basics of genetics, and prepared two 40 min. long expository lectures and supplementary PowerPoint presentations (with around 20 slides each). The teacher also used a whiteboard, especially in the case of genetics, and around 1-2 min. long videos (four for animal learning, two for genetics).

We also picked two relevant micro-games on these topics: *Orbis Pictus Bestialis* (OPB; Brom et al., 2011)[4] and *Bird Breeder* (BB; Novak & Wilensky, 2007). They both feature clearly-defined goals that can be reached within 10 – 20 minutes by the majority of adolescent learners. The BB game also features a score.

[4] The OPB game was originally developed by our research team for educational purposes.

Finally, we prepared supplementary multimedia materials for the teacher to use at appropriate moments during the game-playing session.

4.3.1 Animal Learning – Orbis Pictus Bestialis

The educational objective of the “animal learning” expository lecture plus playing the OPB game was to teach students the basics of ethology, behaviourism and animal learning. The lecture focused on the distinction between ethology and behaviourism; key researchers in these disciplines (Pavlov, Skinner, Lorenz); key research instruments, such as the Skinner box; and key types of learning, such as imprinting, conditioning, habituation and sensitization. The game enabled students to explore major phenomena of positive reinforcement learning, to practice basic animal training techniques (for instance, behaviour capturing, shaping, and chaining) and to learn how to use a clicker during the training.[5]

The slides for the expository lecture included information on the abovementioned concepts, disciplines and persons. We also used four videos on the following topics: a) Konrad Lorenz and imprinting, b) Pavlovian conditioning, c) operant conditioning of a pigeon in a Skinner box, d) demonstrating the behaviour of a trained lemur in a zoo.

In this study, two levels of the OPB game were used. In the first level, learners have to train a dog to wave one front paw based on a verbal stimulus; and in the second level, they have to train a lemur to enter a transportation box and close the door behind it. Each animal is driven by a biologically plausible behavioural model (see Preuss, 2011 for details).

[5] A clicker is a small noisemaker for marking behaviours to be reinforced.

In both levels, animals spontaneously perform certain actions. The player interacts with the animal by presenting it various stimuli (by pressing a button; Fig. 1), such as showing a hand in a specific position, showing a hand holding a certain object, touching the animal, giving it food or pressing a clicker. The learning happens in real time.

The educational goal of the first level is to practice two specific animal training methods: clicker training and shaping (see Brom et al., 2011 for details). The second level is slightly more complex and its goal is to practice another training method: chaining.

The slides to be used by the teacher during the game playing session had primarily screenshots from the game, illustrating key steps from the training procedures. We also used a supplementary video on the course of training a real dog for the in-game task. The same slides and video were used in both conditions.

-- Insert Fig. 1 about here –

4.3.2 Genetics – Bird Breeder

The educational objective of the “genetics” expository lecture was to teach students the basics of Mendelian genetics, the difference between phenotype and genotype, complete and incomplete dominance and co-dominance, the notions of DNA and proteins, the idea behind gene therapy, and the concepts of mitosis and meiosis. The game enabled students, in the role of animal breeder, to breed specific line of animals. Doing so enabled them to practise the rules of Mendelian genetics and co-dominance and to reinforce the notions of genotype vs. phenotype and recessive and dominant alleles.

The slides for the expository lecture included information on the concepts mentioned above. We also used two videos: one on mitosis and one on meiosis. The BB game (Novak & Wilensky, 2007) is a game from a library of educational simulations in the Netlogo toolkit (Wilensky, 1999). It has two levels and we used both.

In the first level, the learner's goal is to breed a specific line of birds. The learner makes decisions based on a simple genetic representation of five traits. Gene expression follows classical Mendelian rules. The learner can either visually inspect the phenotype or analyse the underlying genotype directly (Fig. 2). The game is step-based: i.e. by repeatedly selecting bird couples for mating, the game's goal can be achieved. The goal can typically be achieved in 10 steps or less, but each step requires the student to think carefully about which birds to mate. Unlike in the OPB game, success in this game is rewarded "financially" (Fig. 2, panel "your funds $"). Earnings represent an analogy to score.

The second level differs from the first one in two aspects. First, it features dragons rather than birds (with five different traits). Second, gene expression follows the rules for co-dominance.

The main challenge in the game is that the target animal has to be a recessive homozygote in all but one trait. The expository lecture illustrated the breeding process with one or two traits only: the game presented a new situation for the students.

For the game playing part, we prepared (in both conditions) slides for the teacher. The slides explained the steps in the breeding process and used the in-game graphics. There was no supplementary video, but the teacher frequently used the white-board.

-- Insert Fig. 2 about here –

4.4 Pen-and-Paper Materials

4.4.1 The Knowledge Test

For each topic (genetics, animal learning) we constructed two test versions, assessing the same knowledge but with slightly different questions and in a different order.[6] Students received one version in the immediate testing session and the other in the delayed testing session. The order in which the tests were administered was counter-balanced.

Each test had two groups of open-ended questions: seven questions assessing facts (*factual* questions throughout) and seven (the BB game) or six (the OPB game) assessing deep knowledge (*transfer* questions). It is impossible to separate strictly knowledge that can be acquired in the expository lecture from knowledge that can be acquired/strengthened during the game playing part. For instance, the games instigated additional cues for certain facts presented in the lectures, which could aid in later retrieval of cued information. Nevertheless, the factual questions were more related to what was taught during the expository lectures, rather than during the game playing parts. With the transfer questions it was the other way round. Examples of paired questions are given in Appendix A.

[6] The test was piloted on a sample of participants different from the experimental sample.

Each question could yield up to 1 point. Partially correct answers were rewarded by .25, .5 or .75 point. The tests were graded by two independent evaluators and the results were averaged. The Spearman correlation coefficient between the two scoring persons was high, over.95. The resulting scores' variables were in the range of 0-1. They are indicated as: *Score-f1*, *Score-f2*, *Score-t1*, and *Score-t2* (1: immediate; 2: delayed; f: facts; t: transfer).

The results of the immediate and the delayed tests correlated with each other with medium to large effect size. The Spearman correlations between Score-f1 and f2 were .42 (OPB) and .49 (BB), and between Scores-t1 and t2 were .62 (OPB) and .63 (BB) (see also Table 6, 7 in Appendix C).

4.4.2 Background Data

The background questionnaire solicited age and gender information. We also asked participants the following questions concerning their gaming and ICT experience: "How often do you play computer games?" and "How often do you use computers?" with the scale "*1) less than one hour a week; 2) 1 - 5 hours a week; 3) 6 - 10 hours a week; 4) more than 10 hours*". (*Game play frequency* and *ICT experience* variables throughout).

Students' grades were obtained from the school administration. These consisted of their final grades in the middle of the school year and their expected final grades at the end of the school year (the study's delayed tests were administered two weeks before the end of the school year). An average of these grades was used as a *GradeBio* variable (1 – *the best*; 5 – *the worst*).

4.4.3 Subjective Evaluation

To assess how students liked the whole educational class unit, one question with a 6-point Likert scale (1 – *very good*; 6 – *very bad*) was administered: “How did you like today’s lesson; including playing [the demonstration of] the game?” (*Like* variable). One question investigated self-perception of acquired knowledge: “How much do you think you learnt today?” (1 – *very much*; 6 – *nothing*) (*Learnt* variable).

To solicit information about students’ perception of games’ characteristics in the “I” condition, the following four questions with a 6-point Likert scale were administered: “How do you assess the game based on the following characteristics: a) graphics, b) difficulty of controlling the game, c) entertainment value, d) educational value” (1 – *very good/easy/large*; 6 – *very bad/hard/small*).[7]

4.5 Procedure

The study was conducted as part of regular education one and half months before the end of the school year. Each class participated for one whole school day (beginning: 7:45, 8:00 or 8:35; end: shortly after noon, before lunch). The schedule for a typical day is depicted in Fig. 3. The research team included a male teacher, a male assistant, and – for six of the days – a female independent observer (who was not always present throughout the whole day).

[7] The background questionnaire also included several other questions irrelevant to the present study; mainly open-ended questions on game characteristics, such as “What did you like most about the game and why? _______________”.

Four classes were assigned to the BB condition (two from each school), five to the OPB condition (three from one school and two from the other). Participants were not informed that we used two different treatments. Participants were research-question-blind.

After the introduction, the expository lecture was given to the intact class. Then, after a short break, the class was randomized: half was assigned to the "I" condition and half to the "C" condition (gender balanced). There were never more than 15 students in one half.

Teacher effect has been suspected in literature to be an important confounder in research on the impact of (any) educational technology on learning performance (e.g., Clark, 2012). Therefore, as already said, we used one teacher. That however posed one practical problem: after the randomization, two halves of every class had to be taught parallel one to the other. Instead of recruiting two different teachers, we adopted a shifted schedule (see Fig. 3). This means that after the randomization, one half started with the teacher under either "C" or "I" condition, while the second half was engaged in irrelevant, emotionally neutral activities with the research assistant (in particular, filling in personality tests). Then, after a short break, the two halves switched and the second half received the second condition. The order in which the "C" and "I" conditions were administered in different days was counterbalanced for both topics.

After the randomization, one half of the class relocated – participants assigned to the "I" condition always moved to the computer laboratory. During the game playing period (in both conditions), a PC, a projector, a large projector screen and a whiteboard were available for the teacher to use. In both conditions, the teacher started with a short summary of the expository lecture and by explaining the game's purpose and its goal(s); using the game's first level. The teacher also explained how the game is connected to the topic of the expository lecture. The teacher always ended the game playing period with a short debriefing.

In the “I” condition, each student sat at a separate computer with at least a 17" wide screen. The game was presented to participants as a competition (“be the first/second/third/... to train/breed the animal”), but no tangible rewards were offered since it is known that they can be problematic (Deci, Koestner, & Ryan, 1999; Cameron, Banko, & Pierce, 2001). The students proceeded in a self-paced manner. The teacher moved around the entire classroom. When addressing the whole class (around 3-5 times), he used prepared slides and/or the video. The teacher also made roughly 15-30 visits to individual students, mainly to comment on what happened in the game. At the end (i.e., before the debriefing), around half of the group had usually already finished the game; the rest were stopped.

In the “C” condition, students sat as they would during a regular school lesson. The teacher stood at the front of the class. He repeated around 10 times the scenario of playing the game, commenting on what happened and prompting the class on how to proceed. He also used the prepared slides and/or the video.

The breaks mirrored regular school breaks if possible. After the break following the intervention, participants received the pen-and-paper knowledge test and then the background questionnaire. These materials were filled in anonymously.

Roughly a month after the experiment ended, we visited the schools for the second time. Students received delayed knowledge tests as part of their regular school lesson on a regular school day (these took around 20-25 min.). Students were not informed in advance about the tests. They thus did not have time to prepare for them. Students present at the delayed testing session, but not on the experimental days, were also given the test and they were considered naïve participants (see Footnote (3)).

-- Insert Fig. 3 around here--

4.6 Informal Observations

During six of the experimental days, an independent observer was present for part of the expository lecture and game playing periods and made informal notes on student's overall behavioural activity. The observation was non-invasive; the observer stood aside and did not come near the students.[8]

At the end of each experimental day, the teacher also made informal notes regarding students' overall activity. In the "I" condition, the teacher moved around the entire classroom; therefore, unlike the observer, he had an opportunity to inspect closely what the students were really doing.

4.7 Data Analysis

Since the distribution of analysed data deviated from normal distribution, a robust non-parametric statistics was used preferentially. Non-parametric Wilcoxon rank sum test was used for group comparison: a) comparison of the effect of game presentation ("C" versus "I") on the scores in the knowledge tests and on the affective outcomes, b) comparison of

[8] The primary purpose of the observations was to select students for subsequent focus groups, which were conducted because of a different study (presented in Buchtova et al., 2013).

biographical variables between the “C” and “I” groups, c) comparison of game characteristics between the BB and the OPB games, d) comparison of scores in the delayed knowledge test between students who participated in the experiment and naïve participants. If needed, Holm-Bonfferoni correction for multiple comparisons was used to ensure the appropriate level of significance (Holm, 1979). A group difference with p-adjusted < .1 is reported as a trend and with p-adjusted < .05 as a significant difference. Cliff’s delta was used for estimating the effect size of the difference. The effect size was classified into four categories: negligible (Cliff’s delta <.147), small (Cliff’s delta < .33), medium (Cliff’s delta < .474) and large (Cliff’s delta ≥ .474). This classification corresponds to the commonly used classification of Cohen’s *d*, which represents the effect size for normally distributed data (Romano, Kromrey, Coraggio, & Skowronek, 2006).

In one case, a covariate, GradeBio, had to be taken into account. This was the case where we compared results from the immediate knowledge tests between the “C” and “I” conditions in the BB game. We used a linear model for the comparison. This was possible because the respective data (after the response variables were log-transformed) became similar to normal distribution (as tested by Shapiro-Wilk test).

Spearman correlation was used for evaluation of the association between analysed variables. The effect size of the association was classified into four categories: negligible (|r| < .2), small (|r| < .4), medium (|r|< .6) and large (|r| ≥ .6).

5 Results

5.1 Participants' Characteristics

Table 1 shows the number of students in each group per game and school/class who participated in the experiment and completed both knowledge tests. The averages of the GradeBio, Game play frequency and ICT experience variables (per game and group) are given in Table 2. As concerns the OPB game, we see that we can assume that the groups were sampled equally. However, this cannot be assumed as concerns the BB game and variables GradeBio (due to significant medium effect size difference between the groups) and possibly also ICT experience (non-significant small effect size difference).

-- Insert Tab. 1 about here –

-- Insert Tab. 2 about here --

5.2 Game Characteristics

Table 3 shows how "I" condition students rated the games according to Graphical quality, Difficulty in controlling the game, Entertainment value and Educational value. Generally, the values (except for the BB's graphics) are low, which means that the games tended to be perceived positively and as easy to master (which is a manipulation check).

We see that the games were comparable as concerns the last three characteristics. The games differ in subjectively-assessed graphical quality; which is not surprising (cf. Fig. 1 and 2). Frequent gamers also probably did not like the OPB game, unlike the BB game (see Appendix B).

-- Insert Tab. 3 about here --

5.3 Learning Outcomes

The results of the knowledge tests are presented in Table 4. For the sake of completeness, correlation matrices of the key variables are given in Appendix C.

-- Insert Tab. 4 about here –

-- Insert Tab. 5 about here --

-- Insert Tab. 6 about here --

As concerns the OPB game, we found negligible between-group differences in factual questions; both for the immediate and delayed test. We found small differences in transfer questions in favour of the “C” condition, but these differences were non-significant. The between-group differences in knowledge decrements were negligible.

As concerns the BB game, “I” group participants tended to score comparably higher on factual questions immediately after the lesson (i.e., in Score-f1; medium effect size) but not one month later (negligible effect size). The “I” condition also performed higher compared to the “C” condition on transfer questions, but the differences ranged from negligible to small and they were non-significant. The between-group differences in knowledge decrements were small in favour of the “C” condition, but again non-significant.

As concerns the BB game, the biology grade was positively correlated to the test scores (see Appendix C). Because the groups were not sampled equally for this game with respect to the GradeBio variable, the between-group difference in Score-f1 for the BB game could be partly due to this unequal distribution. Similarly, the unequal distribution could party contribute to the non-significant between-group differences in Score-t1, -f2, and -t2 (i.e., the “C” group could score better if there were students with as good biology grades in the “C” group as in the “I” group). As a remedy, we chose a statistical approach in which the scores from knowledge tests were predicted by a linear model with factor Group and covariate GradeBio. The distribution of scores from the knowledge tests in the BB game alone became similar to normal distribution after log-transformation, as tested by Shapiro-Wilk test.

The linear model was applied separately to both tests and both types of questions. In the case of the immediate test, the factor Group and the covariate GradeBio had significant effect on the response variable Score-f1 (Group: *F*(1, 67) = 5.528, *p* = .022; GradeBio: *F*(1, 67) = 8.255, *p* = .006) but not on the response variable Score-t1 (Group: *F*(1, 67) = 2.217; *p* = .141; GradeBio: *F*(1, 67) = 0.774, *p* = .382).[9] In the case of the delayed test, neither the factor Group nor the covariate GradeBio had significant effect on the response variable Score-f2 (Group: *F*(1, 67) = 1.155, *p* = .29, GradeBio: *F*(1, 67) = 1.911, *p* = .17). The covariate had an effect on the response variable Score-t2 (GradeBio: *F*(1, 67) = 4.757, *p* < .033), but no effect of Group on Score-t2 was found (*F*(1, 67) = 0.340, *p* = .56). This indicates that the between-group difference for Score-f1 still holds after the effect of GradeBio is accounted for. At the same time, no new between-group differences in test scores is revealed when unequal distribution with respect to biological grades is considered.

As concerns the BB game, the groups were probably also not sampled equally with respect to the ICT experience variable. However, this variable was uncorrelated with test scores for this game (Appendix C). Therefore, it is unlikely that this biographical variable affected the test scores.

5.4 Affective Outcomes

Between-condition differences for the Learnt variable were negligible for both games (Tab. 5). As concerns the Like variable, small (but approaching medium range) significant difference was found for the BB game but not for the OPB game; in favour of the “I”

[9] Note that grades in biology were not available to us in the case of 3 students (for the BB game).

condition (Tab. 5). The relationship between these two variables and learning outcomes is described in Appendices C and D.

-- Insert Tab. 5 about here –

5.5 Informal Observations

Concerning the game playing periods, the observer’s and the teacher’s informal observations provided the following findings.

The “C” condition: when the teacher asked students on how to proceed in the game, he usually got 1-3 independent answers by different students. A brief discussion among a few students was also sometimes triggered (usually among up to 3-5 students; only in one class did this happen regularly among half of the “C” group). It was less often that no-one responded to the teacher. In general, about half of the “C” group in every class did not seem to be engaged much by the teacher playing the game.

The “I” condition: at least one-fourth of class usually clicked their way through the game (without elaborating on what is on the screen); expressed a desire to switch on their email/Facebook; or was engaged in irrelevant activities within the game (such as arranging dragons on the screen to make a specific pattern or attempting to overfeed the dog). On the other hand, some students (usually one or two), but in two groups around half of the class, wanted to continue playing the game during the break; especially in the case of the BB game. It was particularly problematic for the teacher to switch to the second level (in both games), because students proceeded at different paces and those who were still deep in the first level when the teacher started explaining the second one, became distracted.

6 Discussion and Conclusion

The topic of this explorative study was the effective use of digital, educational micro-games in classrooms. We asked the question of whether an educational method capitalizing on collective game play is more, equally or less effective compared to a method capitalizing on individual play. Both conditions used the same micro-game, whose purpose was to reinforce and integrate part of the knowledge learnt in the introductory lecture. In both conditions, classes of up to 15 students were engaged. In both conditions, the teacher showed supplementary multimedia materials via a projector and provided short verbal inputs during the game playing section. We used two different micro-games (OPB and BB) with similar complexity.

The main result is that the “C” condition is not much worse than the “I” condition in terms of learning achievements and subjective evaluation of the educational unit. The between-condition knowledge gain differences for students who played the OPB game were negligible to small (and non-significant): both for immediate and delayed knowledge tests and both for transfer and factual questions (Tab. 4). The pattern was similar for the BB game with one exception: in immediate achievement in factual questions, the “I” group participants outperformed the “C” group participants (medium effect size; marginally significant; Tab. 4). This difference could not be explained by the comparably higher biological grades for the “I” group participants (Sec. 5.3). As concerns self-estimated learning, we found no between-condition differences (Tab. 5). A significant difference in favour of the “I” condition was found for the BB game (but not the OPB game) in enjoyment (small to medium effect size; Tab. 5).

In summary, individual play was favoured only by one out of eight between-group learning achievement differences and by one out of four between-group differences in affective variables (while no case clearly favoured the collective condition). Given that it is easier and cheaper to play the game collectively rather than individually in schools (as discussed in Sec. 1), this study's main practical implication is that collective play is a method that should be seriously considered by educational designers and teachers.

These findings can also be easily explained from the perspective of the active learning view (Mayer, 2009) and the INTERACT model (Domagk et al., 2010). Collective play in a classroom probably provoked active cognitive participation for a certain number of participants. Individual play in a lab probably provoked active cognitive participation for a slightly higher number of participants. However, this advantage of the individual play was probably counterbalanced by the teacher's interruptions of the game play that could cause distraction and/or cognitive overload (see Sec. 2). This interpretation is also supported by the informal observations (Sec. 5.5): about half of students were engaged by the collective play, less than half were not engaged by the individual play and some were distracted by the teacher's inputs in the individual play. In the future, it would be useful to pay attention to methods of promoting *cognitive* activity in game-based learning in schools and/or reducing learners' cognitive distraction. For instance, would two teachers in our "I" condition resolve the issue with distraction? Can problems with asynchronous play be resolved when students do not play individually but in groups of 2-4 persons per computer (cf. Watson et al., 2011)?

Because the OPB and BB games do not focus primarily on teaching facts, it is somewhat surprising that the only knowledge gain difference was found for factual questions rather than transfer questions (and only for the BB game). However, this can be explained. First, as already said, factual questions were not related only to knowledge that could be acquired from the expository lecture (Sec. 4.4.1). The games could instigate cues for certain facts presented in the lecture, which could aid in later retrieval. Because the BB game was more symbolic than the OPB game (it presented the information to students in symbols: the animals' visible traits and their genotype), it could present these cues (for instance, as concerns the difference between haploid and diploid gametes or between genes and alleles) more easily compared to the other game. Second, these cues could also be more accessible to the learners in the "I" condition, because they were better visible from the monitor compared to the projector screen (they were quite small: see Fig. 2). This could even be reflected in the lower enjoyment the "C" condition participants had for the BB game: perhaps they did not see all the details on the projector screen and thus liked the educational experience less. Third, because this difference was apparent in immediate tests only, learners' cognitive participation was probably sufficient to utilize the added value of the cues in the initial information processing, but not enough to boost retention over a month-long period. Fourth, these cues could be more beneficial for retrieving superficially acquired information from the expository lecture, i.e., facts, rather than for acquiring new conceptual knowledge from the game playing by deep learning (otherwise, we would also see the effect in transfer tests).

However, given that this was only an exploratory study, one must be cautious in the interpretation of these findings. We did not measure cognitive distraction and/or behavioural/cognitive activity and/or cognitive load (this is in fact quite difficult; see Brünken, Plass, & Leutner, 2003). Therefore, all the explanations above should be considered as hypotheses for future studies rather than firm conclusions. It would be useful to collect more quantitative information on the level of participants' cognitive activity and distraction and also to investigate if games like Bird Breeder can really present cognitive cues facilitating retrieval of factual information from an expository lecture. This study is also not without limitations. First, it would be useful to have pre-tests. In our case, they would have to be administered at least several days before the study (to avoid cueing as to what should be remembered) and that was unfortunately not possible due to time constraints. Nevertheless, we are confident that our participants improved well beyond the baseline (see Footnote 3).

Second, our shifted schedule could create some differences between the two groups participating during one day (e.g., due to the teacher fatigue). The groups' order was counterbalanced, but with only 4 or 5 classes for one topic/game, there is still a risk of confounding the research outcome. However, the only alternative was to have two different teachers. In our opinion, this risk is smaller in the case of the shifted schedule compared to having two different teachers.

Third, even though our findings concerning both micro-games are superficially similar (which would point towards generalizability at least for micro-games that have similar complexity and which are used in a similar way to the one we used), we found a difference in immediate factual tests mentioned above and also in that frequent gamers probably did not like the OPB game, which was not the case of the BB game (Appendix B). This indicates that even two apparently similar micro-games can have a different appeal to different audiences and possibly also different impacts on learning processes. Understanding the reasons behind these differences is an important topic for future research.

To conclude, in our opinion, the limitations of this study do not undermine its main conclusion. This conclusion is that the method of collective play should be considered in future research and practice, because it may be cost-effective for some games compared to individual play.

Appendix A – Examples of Test Questions

Examples of paired questions are:

- Genetics (factual questions):
    - Which plant did Gregor Johann Mendel do most of his research on and what did he discover?
    - In what century did Gregor Johann Mendel live and what was he famous for?
- Genetics (transfer questions):
    - If we cross a red-and-white-haired shorthorn bull and cow, a red-and-white-haired calf might be born. However (less often), the calf might be completely white or completely red. What colour would the parent cattle have to have in order for us to be sure that the calf would be red-and-white-haired? Explain why your proposal should work.
    - Natalie crossed a pink rose with another pink rose. From the seeds grew roughly half pink roses and the other half was red and white. What roses would Natalie have to cross to ensure that only pink roses grew from the seeds? Explain why your proposal should work.
- Animal training (factual questions):
    - Describe how operant conditioning works.
    - What is the difference between operant and classical conditioning?
- Animal training (transfer questions):
    - Thomas' ferret can wag its tail at the sound of a whistle. Thomas would now like to teach it to do so based on the command: "Do it!". He's trying to do so by first whistling and then saying "do it" and then he rewards the ferret, if it

does the trick. However, when he just uses the command “do it”, the ferret will not wag its tail at all. Has Thomas made a mistake somewhere? Explain.

- Johnathan’s dog can roll over based on a clap command. Johnathan would now like to teach his dog to do so based on the verbal command: “roll over”. He attempts to do so by clapping, then saying “roll over” and then ultimately rewarding the dog, when it does the trick correctly. Yet when Johnathan tries just command “roll over”, the dog refuses to do so. Has Johnathan made a mistake somewhere? Explain.

Appendix B – Differences between the Games

We found that frequent gamers probably did not like the OPB game; unlike the BB game. There was a difference between the games in the relationship between Entertainment value and Game play frequency (BB: *r* = -.31; *p* = .06; OPB: *r* =.47; *p* =.002; because Entertainment value and Game play frequency have opposite scales, a negative correlation means a positive relationship) and between Educational value and Game play frequency (BB: *r* = -.04; *p* = .80; OPB: *r* = .38; *p* = .014). We also saw a between-game difference in immediate factual tests (Sec. 5.3) accompanied by a difference in enjoyment (Sec. 5.4).

These differences deserve some attention because some researchers in the field of digital-game-based learning intuitively assume that similar games produce similar learning and affective outcomes. It would be useful in the future to investigate what game elements (and in what socio-spatial contexts) are useful for which types of learners. In our opinion, attention should be paid to structural features, such as the utilizing of symbolic representation in the BB game, rather than superficial features, such as instructional domain (e.g., biology).

Appendix C – Correlations between Variables

-- Insert Tab. 6 about here –

-- Insert Tab. 7 about here --

Appendix D – The Relationship between Affective and Cognitive Variables

Recently, the field of multimedia learning has become interested in investigating the relationship between affective and cognitive outcomes in intervention studies (e.g., Brom et al., 2014; Plass et al., 2014). Even though investigation of this relationship is out of the primary scope of this study, we report on the main correlations.

There is a large inter-correlation between the Like and the Learn variable (Tab. 6, 7), indicating a common denominator. Correlation between test scores and these two variables are generally positive, but mostly non-significant and in the negligible to small range (Tab. 6, 7). These non-significant correlations are not that surprising for two reasons. First, our measures of affective variables are rather crude (cf. Plass et al., 2013; Wang et al., 2008). If affectivity is in the primary scope of this work, other measures, such as in situ administered Flow Short Scale (Rheinberg, Vollmeyer, & Engeser, 2003) or PANAS (Watson, Clark, & Tellegen, 1988) would be more appropriate. Second, as argued in Sec. 2, higher positive affect may not always imply higher learning gains (because of a possible distraction). Therefore, future multimedia learning studies focusing on the affectivity–learning link should also consider measuring distraction/cognitive load (which is not an easy task). In this particular study, the relationship between affect and learning is probably complex in the "individual" and "collective" conditions. For instance, social interaction anxious persons might prefer the individual play, in which they can better avoid interaction with their peers; whereas, extroverts may prefer collective play. Therefore, the interplay between social contexts and personal characteristics should be also considered in future research.

There is also a relationship between three game characteristics, Graphics, Entertainment Value or Educational Value, and Learnt/Like variables (predominantly medium effect sizes; BB: *r* in the range from .28 to .69; *p* < .1; OPB: *r* in the range from .22

to .46; $p < .15$). This relationship is weaker or absent for the Difficulty of Controlling the Game variable (OPB: Control difficulty x Like: $r = .3$, $p < .1$; Control difficulty x Learnt: $r = .24$, $p = .13$; BB: $r < .09$; $p > .57$), which is a meaningful outcome. There is almost no relationship between game characteristics and test scores: of 32 correlations, two approach significance and two are significant, which is an outcome probably caused by chance alone.

References

All, A., Nunez Castellar, E. P., & Van Looy, J. (2014). Measuring Effectiveness in Digital Game-Based Learning: A Methodological Review. *International Journal of Serious Games, 1*(2), 3-21.

Baddeley, A. D., Eysenck, M., & Anderson, M. C. (2009). *Memory*: Hove: Psychology Press.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). *Off-task behavior in the cognitive tutor classroom: when students game the system*, in the Proceedings of the SIGCHI conference on Human factors in computing systems.

Bourgonjon, J., Valcke, M., Soetaert, R., De Wever, B., & Schellens, T. (2011). Parental acceptance of digital game-based learning. *Computers & Education, 57*(1), 1434-1444.

Bourgonjon, J., Valcke, M., Soetaert, R., & Schellens, T. (2010). Students' perceptions about the use of video games in the classroom. *Computers & Education, 54*(4), 1145-1156.

Brom, C., Buchtová, M., Šisler, V., Děchtěrenko, F., Palme, R., & Glenk, L. M. (2014). Flow, Social–Interaction anxiety and salivary cortisol responses in serious games: A Quasi-experimental study. *Computers & Education, 79*, 69-100.

Brom, C., Preuss, M., & Klement, D. (2011). Are educational computer micro-games engaging and effective for knowledge acquisition at high-schools? A quasi-experimental study. *Computers & Education, 57*(3), 1971-1988.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 53-61.

Buchtová, M., Šisler, V., Brom, C. (2013). Educational Games and Simulations at School: Experimental Comparison with Classic Teaching Methods and Requirements of Successful Implementation into School Environment and Curricula. *Proceedings of* *the 2nd International Workshop on Interaction Design in Educational Environments* (pp. 125-132) SciTePress

Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *The Behavior Analyst, 24*(1), 1-44.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational psychology review, 3*(3), 149-210.

Clark, R. E. (Ed.). (2012). *Learning from Media: Arguments, Analysis, and Evidence, Second Edition* (2nd. ed.): Information Age Publishing.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin, 125*(6), 627-668.

De Grove, F., Bourgonjon, J., & Van Looy, J. (2012). Digital games in the classroom? A contextual approach to teachers' adoption intention of digital games in formal education. *Computers in Human Behavior, 28*(6), 2023-2033.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of educational psychology, 100*(1), 223-234.

Domagk, S., Schwartz, R. N., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human behavior, 26*(5), 1024-1033.

Egenfeldt-Nielsen, S. (2005). *Beyond Edutainment: Exploring the Educational Potential of Computer Games.* (Ph.D. thesis), University of Copenhagen.

Gjedde, L. (2013). Role Game Playing as a Platform for Creative and Collaborative Learning *Proceedings of European Conference on Games Based Learning '13* (pp. 190-197).

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65-70.

Huizenga, J., Admiraal, W., & Ten Dam, G. (2013). *Teaching with games in secondary education in the Netherlands.* Paper presented at the EARLI 2013, Munich, http://www.earli2013.org/programme/proposal-view/?abstractid=3426. Accessed 25 Sept 2014.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational researcher, 38*(5), 365-379.

Judd, C. M., Smith, E. R. and Kidder, L. H. (1991) *Research Methods in Social Relations,* Fort Worth: Hartcourt Brace.

Kebritchi, M. (2010). Factors affecting teachers' adoption of educational computer games: A case study. *British Journal of Educational Technology, 41*(2), 256-270.

Kenny, R. F., & McDaniel, R. (2011). The role teachers' expectations and value assessments of video games play in their adopting and integrating them into their classrooms. *British Journal of Educational Technology, 42*(2), 197-213.

Ketelhut, D. J., & Schifter, C. C. (2011). Teachers and game-based learning: Improving understanding of how to increase efficacy of adoption. *Computers & Education, 56*(2), 539-546.

Klopfer, E. (2008). *Augmented learning: Research and design of mobile educational games*: MIT Press.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*(1), 14-19.

Mayer, R. E. (2009). *Multimedia Learning* (2nd ed.): Cambridge University Press.

Moreno, R. (2005). Instructional technology: Promise and pitfalls. *Technology-based education: Bringing researchers and practitioners together* (pp. 1-19): Information Age Publishing.

Moreno, R. (2010). Cognitive load theory: More food for thought. *Instructional Science, 38*(2), 135-141.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational psychology review*, *19*(3), 309-326.

Novak, M. and Wilensky, U. (2007). NetLogo Bird Breeder model. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, available: http://ccl.northwestern.edu/netlogo/models/BirdBreeder [accessed 25 Sept 2014].

Plass, J. L., Heidig, S., Hayward, E. O., Homer, B. D., & Um, E. (2014). Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction, 29*, 128-140. doi: 10.1016/j.learninstruc.2013.02.006

Plass, J. L., O'Keefe, P. A., Homer, B. D., Case, J., Hayward, E. O., Stein, M., & Perlin, K. (2013). The Impact of Individual, Competitive, and Collaborative Mathematics Game Play on Learning, Performance, and Motivation. *Journal of educational psychology, 105*(4), 1050-1066. doi: 10.1037/a0032688

Proctor, M. D., & Marks, Y. (2013). A survey of exemplar teachers' perceptions, use, and access of computer-based games and technology for classroom instruction. *Computers & Education, 62*, 171-180.

Preuss, M. (2011). (2011). *A Computational Model of an Animal Designed for High-School Education.* (MSc thesis), Charles University in Prague [in Czech].

Rheinberg, F., Vollmeyer, R., & Engeser, S. (2003). Die Erfassung des Flow-Erlebens [in German]. In J. Steinsmeier-Pelster & F. Rheinberg (Eds.), *Diagnostik von Motivation und Selbstkonzept* (pp. 261-279): Hogrefe. [in German]

Romano, J., Kromrey, J. D., Coraggio, J. and Skowronek, J. (2006) *Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NNSE and other surveys?*, Paper presented at Annual meeting of the Florida Association of Institutional Research, Florida.

Sandford, R., Ulicsak, M., Facer, K., Rudd, T. (2006) *Teaching with Games: Using commercial off-the-shelf computer games in formal education*, Bristol: Futurelab, available: https://www.nfer.ac.uk/publications/FUTL49/FUTL49.pdf [accessed 20.2. 2015].

Sisler, V., & Brom, C. (2008). Designing an educational game: Case study of 'Europe 2045'. In Z. Pan, A. D. Cheok & W. Muller (Eds.), *Transactions on Edutainment I* (Vol. 5080, pp. 1-16).

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64*(2), 489-528.

Sweller, J. (1999). *Instructional design in technical areas.* Camberwell, Australia: ACER Press.

Szewkis, E., Nussbaum, M., Rosen, T., Abalos, J., Denardin, F., Caballero, D., . . . Alcoholado, C. (2011). Collaboration within large groups in the classroom. *International Journal of Computer-Supported Collaborative Learning, 6*(4), 561-575.

Thomas, R., & Hooper, E. (1991). Simulations: An opportunity we are missing. *Journal of research on computing in education, 23*(4), 497-513.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*(3), 229-243.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies, 66*(2), 98-112.

Wastiau, P., Kearney, C., & den Berge, W. V. (2009). *How are digital games used in schools? Complete Results of the Study.*: European Schoolnet. (Accessed 27 Sept 2014) http://games.eun.org/upload/gis-full_report_en.pdf.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology, 54*(6), 1063-1070.

Watson, W. R., Mong, C. J., & Harris, C. A. (2011). A case study of the in-class use of a video game for teaching high school history. *Computers & Education, 56*(2), 466-474.

Wilensky, U. (1999) NetLogo, Center for Connected Learning and Computer-Based Modeling, Northwestern University, available: http://ccl.northwestern.edu/netlogo/ [accessed 25 Sept 2014].

Williamson, B. (2009). *Computer games, schools, and young people: A report for educators on using games for learning*: Bristol: Futurelab.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games. *Journal of educational psychology*, *105*(2), 249-265. doi: 10.1037/a0031311

Figures

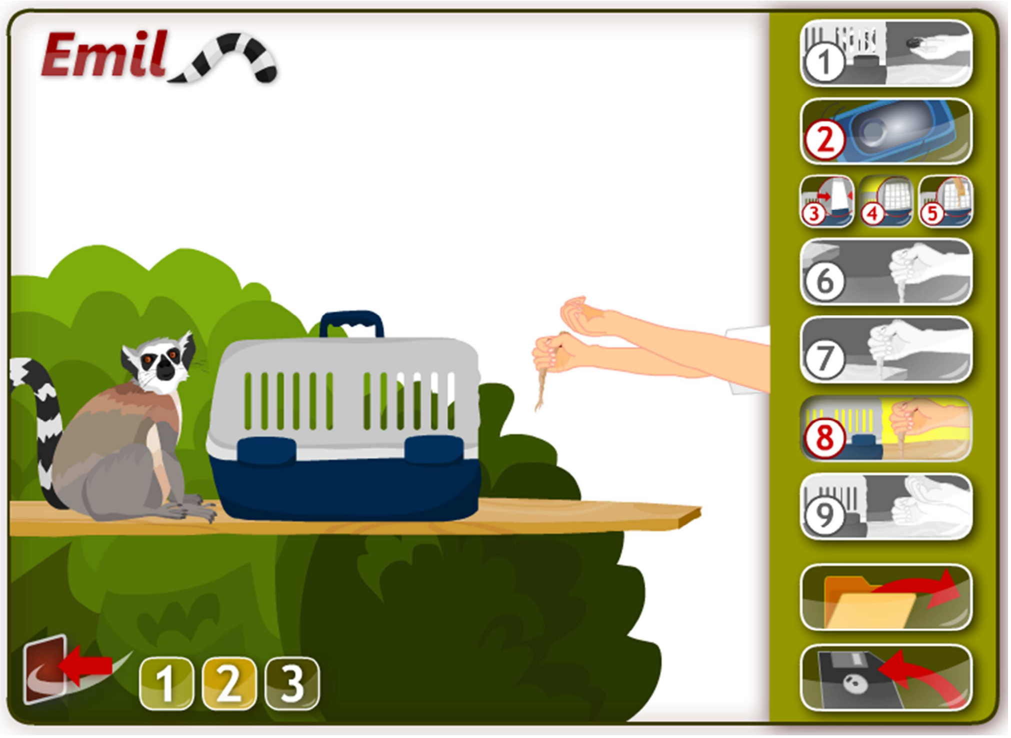
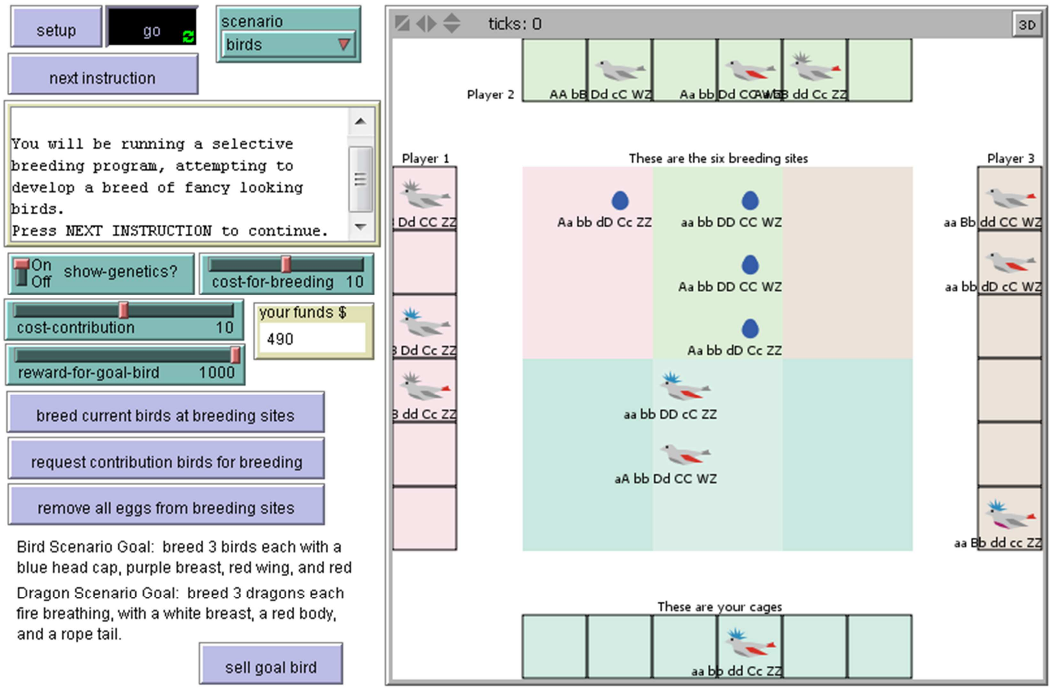


Figure 1. A screenshot from the OPB game: training the lemur, Emil, to enter the transportation box. Control buttons are located on the right side of the figure. In the training step depicted, the trainer holds a rope in one hand, while making a visual stimulus with the other hand (the rope will be hanged on a pulley fastened to the box and it will close the door when pulled by the lemur – in consecutive steps).



scenario: birds

ticks: 0

You will be running a selective breeding program, attempting to develop a breed of fancy looking birds.
Press NEXT INSTRUCTION to continue.

cost-for-breeding 10

cost-contribution 10

reward-for-goal-bird 1000

your funds $ 490

Bird Scenario Goal: breed 3 birds each with a blue head cap, purple breast, red wing, and red

Dragon Scenario Goal: breed 3 dragons each fire breathing, with a white breast, a red body, and a rope tail.

Figure 2. GUI: Bird Breeder, 1st level, bird scenario. The bird’s genotype is shown below each bird. We used a localized version in the experiment. (with permission of Uri Wilensky)

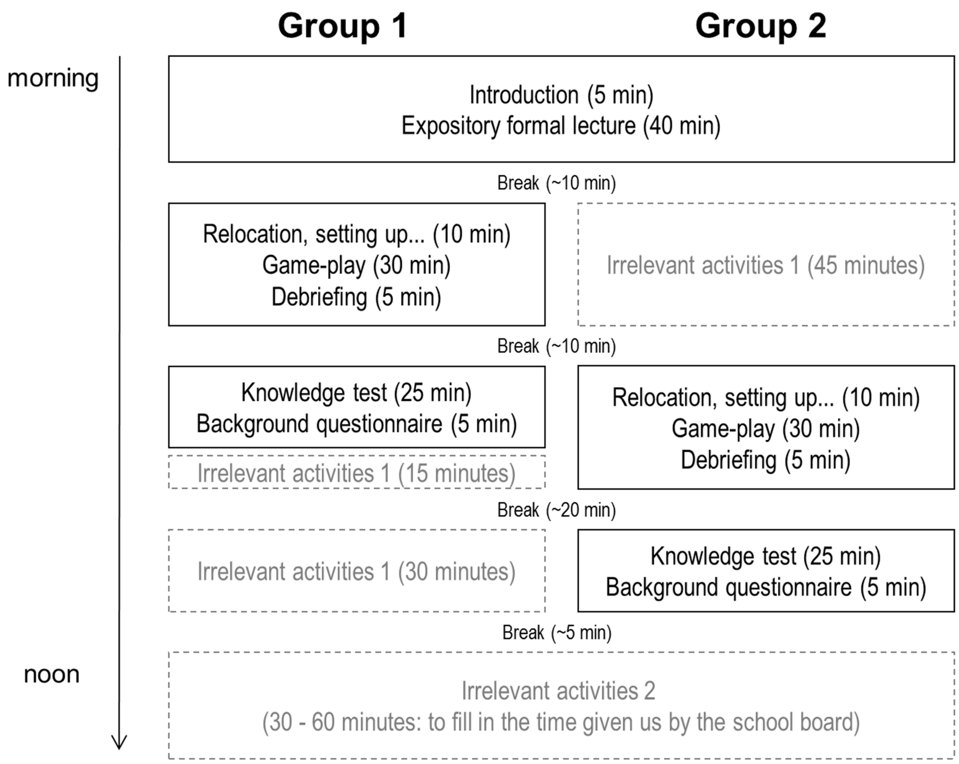

Group 1
Group 2
morning
noon
Introduction (5 min)
Expository formal lecture (40 min)
Break (~10 min)
Relocation, setting up... (10 min)
Game-play (30 min)
Debriefing (5 min)
Irrelevant activities 1 (45 minutes)
Break (~10 min)
Knowledge test (25 min)
Background questionnaire (5 min)
Irrelevant activities 1 (15 minutes)
Relocation, setting up... (10 min)
Game-play (30 min)
Debriefing (5 min)
Break (~20 min)
Irrelevant activities 1 (30 minutes)
Knowledge test (25 min)
Background questionnaire (5 min)
Break (~5 min)
Irrelevant activities 2
(30 - 60 minutes: to fill in the time given us by the school board)

Figure 3. Schedule of an experimental day.

Tables

Table 1

Number of Students Included in the Analysis

		Condition		
Game	Class	"C"	"I"	"C" + "I"
OPB	A	13	9	22
	B	5	7	12
	C	8	9	17
	D	12	10	22
	E	13	7	20
	All (A – E)	51	42	93
Bird breeder	F	9	11	20
	G	9	7	16
	H	12	11	23
	I	6	8	14
	All (F – I)	36	37	73

Table 2

Participants' Background Characteristics

Game	Treatment		Statistics	GradeBio	ICT experience	Frequency game playing
OPB	"C"		Median	2.5	3	1
			Mean	2.367	2.902	1.529
			SD	0.769	0.755	0.809
	"I"		Median	2.5	3	1
			Mean	2.494	3.024	1.833
			SD	0.827	0.811	1.102
		Effect size	Cliff's delta	.091	.077	.131
			effect	negligible	negligible	negligible
		Wilcoxon test	*W*	935	988.5	931
			p	.448	.5	.224
			p (adjusted)	.896	.896	.672
Bird breeder	"C"		Median	2.5	3	1
			Mean	2.515	2.771	1.528
			SD	0.691	0.91	0.91
	"I"		Median	2	3	1
			Mean	2.021	3.135	1.703
			SD	0.713	0.855	0.968
		Effect size	Cliff's delta	- .382	.239	.116
			effect	medium	small	negligible
		Wilcoxon test	*W*	845.5	493	589
			p	.006	.068	.327
			p (adjusted)	.018*	.136	.327

Note: A higher value means a worse grade but a higher ICT experience/more frequent playing.

**p* < .05

Table 3

Participants' Subjective Evaluation of the Games (for the "I" condition)

Game		Statistics	Graphical quality	Difficulty in controlling the game	Entertainment value	Educational value
OPB		Median	3	2	2.5	2
		Mean	2.762	1.786	2.81	2.405
		SD	1.031	1.048	1.366	1.083
Bird breeder		Median	3	1	2	2
		Mean	3.703	1.649	2.324	2.27
		SD	1.199	0.978	1.056	1.289
	Effect size	Cliff's delta	.449	- .076	- .183	- .127
		effect	medium	negligible	small	negligible
	Wilcoxon test	*W*	428	836	919.5	875.5
		p	.0003	.526	.148	.314
		p (adjusted)	.0012**	.628	.444	.628

Note: "1" is the best possible score; "6" is the worst possible score.

***p* < .01

Table 4

Test Scores and Between-group Differences

				Immediate test		Delayed test		Immediate – Delayed[a]	
Game	Treatment		Statistics	Facts	Transfer	Facts	Transfer	Facts	Transfer
OPB	“C”		Median	0.25	0.458	0.089	0.375	0.107	0.063
			Mean	0.234	0.423	0.11	0.342	0.124	0.081
			SD	0.146	0.177	0.117	0.215	0.128	0.159
	“I”		Median	0.214	0.375	0.089	0.219	0.107	0.094
			Mean	0.234	0.363	0.114	0.28	0.12	0.083
			SD	0.159	0.148	0.121	0.186	0.127	0.165
		Effect size	Cliff’s delta	- .047	- .208	.033	- .178	- .019	.003
			effect	negligible	small	negligible	small	negligible	negligible
		Wilcoxon test	*W*	1121	1293.5	1035.5	1262	1091.5	1068
			p	.702	.086	.785	.141	.877	.985
			p (adjusted)	1	.516	1	.705	1	1
Bird breeder	“C”		Median	0.402	0.286	0.357	0.196	0.089	0.062
			Mean	0.439	0.293	0.352	0.213	0.087	0.079
			SD	0.143	0.173	0.136	0.138	0.135	0.144
	“I”		Median	0.554	0.339	0.375	0.196	0.125	0.089
			Mean	0.53	0.353	0.38	0.237	0.15	0.116
			SD	0.161	0.171	0.117	0.163	0.149	0.126
		Effect size	Cliff’s delta	.333	.183	.105	.061	.22	.203
			effect	medium	small	negligible	negligible	small	small
		Wilcoxon test	*W*	444.5	544	596	625.5	519.5	530.5
			p	.015	.180	.443	.659	.107	.136
			p (adjusted)	.090†	.544	.886	.886	.535	.544

[a] The difference is computed for each participant as Score-1 – Score-2.

†*p* < .10

Table 5

Self-estimated learning (Learnt variable) and Enjoyment of the lesson (Like variable)

Game	Treatment		Statistics	Like	Learnt
OPB	“C”		Median	2	2.5
			Mean	2.294	2.627
			SD	0.986	1.22
	“I”		Median	2	2
			Mean	2.095	2.333
			SD	0.759	0.846
		Effect size	Cliff’s delta	- .087	- .127
			effect	negligible	negligible
		Wilcoxon test	*W*	1164	1207
			p	.433	.269
			p (adjusted)	.538	.538
Bird breeder	“C”		Median	2	2
			Mean	2.556	2.5
			SD	0.998	1.108
	“I”		Median	2	2
			Mean	2.014	2.351
			SD	0.87	0.949
		Effect size	Cliff’s delta	- .321	- .065
			effect	small	negligible
		Wilcoxon test	*W*	880	709
			p	.011	.617
			p (adjusted)	.022*	.617

*$p < .05$

Note: “1” is the best possible evaluation; “6” is the worst possible evaluation.

Table 6

Correlation Matrix for the OPB game (Spearman correlation coefficient).

	2.	3.	4.	5.	6.	7.	8.	9.
1. Score-f1	.42***/m	.44***/m	.39***/s	-.25*/s	-.06_/n	-.23*/s	-.08_/n	-.08_/n
2. Score-f2	–	.43***/m	.55***/m	-.34**/s	.14_/n	-.04_/n	-.14_/n	-.2†/s
3. Score-t1		–	.62***/l	-.17_/n	.03_/n	-.1_/n	-.06_/n	-.09_/n
4. Score-t2			–	-.26*/s	.18†/n	-.19†/n	-.13_/n	-.09_/n
5. GradeBio				–	.18†/n	.21*/s	-.02_/n	.2†/s
6. ICT experience					–	.41***/m	-.05_/n	-.13_/n
7. Freq. gameplay						–	.03_/n	-.14_/n
8. Like							–	.6***/l
9. Learnt								–

_p > .10 †p < .10 *p < .05 **p < .01 ***p < .001

Effect sizes: n = negligible; s = small; m = medium; l = large

Note: P-values were not adjusted for multiple comparisons; therefore, some correlations are likely spurious. For Score-xy variables, ICT experience and Frequency of game playing, the higher values mean higher scores/experience/frequency. For GradeBio, Like and Learnt variables, the lower values mean the better grade/evaluation.

Table 7

Correlation Matrix for the BB game (Spearman correlation coefficient).

	2.	3.	4.	5.	6.	7.	8.	9.
1. Score-f1	.49***/m	.41***/m	.22†/s	-.35**/s	.07_/n	.01_/n	-.28*/s	-.19†/n
2. Score-f2	–	.47***/m	.31**/s	-.22†/s	.07_/n	.05_/n	-.23†/s	-.23*/s
3. Score-t1		–	.63***/l	-.18_/n	-.05_/n	.14_/n	-.03_/n	-.19_/n
4. Score-t2			–	-.28*/s	-.07_/n	0_/n	-.13_/n	-.29*/s
5. GradeBio				–	.1_/n	.13_/n	.1_/n	.05_/n
6. ICT experience					–	.59***/m	-.13_/n	.04_/n
7. Freq. gameplay						–	-.03_/n	.02_/n
8. Like							–	.65***/l
9. Learnt								–

_p > .10 †p < .10 *p < .05 **p < .01 ***p < .001

Effect sizes: n = negligible; s = small; m = medium; l = large

Note: P-values were not adjusted for multiple comparisons; therefore, some correlations are likely spurious. For Score-xy variables, ICT experience and Frequency of game playing, the higher values mean higher scores/experience/frequency. For GradeBio, Like and Learnt variables, the lower values mean the better grade/evaluation.