Turning High-Schools into Laboratories? Lessons Learnt from Studies of Instructional Effectiveness of Digital Games in the Curricular Schooling System

[1]Cyril Brom, [1,2]Vít Šisler, [1,2]Michaela Buchtová, [3]Daniel Klement, [3]David Levčík

[1]Charles University in Prague, Faculty of Mathematics and Physics
Ke Karlovu 3, Prague, Czech Republic
[2]Charles University in Prague, Faculty of Arts,
Nám. Jana Palacha 2, Prague, Czech Republic
[3]Institute of Physiology, Academy of Sciences of the Czech Republic
Vídeňská 1083, 142 00, Prague, Czech Republic

Abstract. Digital games are believed by many to be instructionally effective in the context of the formal schooling system; however, studies investigating this idea empirically are limited and their outcomes are often inconclusive and/or difficult to interpret. Part of the problem is caused by the fact that when conducting a study in an authentic environment, i.e., in a school, as opposed to a laboratory, researchers encounter many common, yet unspoken, technical pitfalls. This paper verbalizes some of these pitfalls and organizes them into 6 Recommendations for "best practice" in field studies on the instructional effectiveness of digital game-based learning (DGBL). These recommendations are based on experience gained during five DGBL studies on more than 700 subjects in the context of secondary education and can be useful to other researchers willing to run similar studies.

Keywords: digital game-based learning, serious games, simulations, secondary schools, instructional effectiveness, empirical studies, learning effects

1 Introduction

In the past decade, many scholars have argued that serious games present a new instructional technology with many potential advantages in the context of the formal schooling system, e.g. [2, 10, 11]. Nowadays, serious games gradually enter schools [27]. However, at the same time, data supporting the idea of the instructional effectiveness of digital game based learning (DGBL) are still limited, inconclusive and/or difficult to interpret, e.g. [5, 12, 20, see also 24]. Coincidentally, in the neighboring field of educational simulations (including non-computer based simulations), Feinstein and Cannon complained in 2002 that claims about the benefits of simulations have remained inconclusive since the Sixties [9; cf. 29]. Moreover, Moreno also noted that other "technological innovations in the past have been considered promising media to promote learning," [21, p. 2], including motion pictures, radio and television [7; cited from 21, p. 2]; however, the "hopes and expectations were largely unmet" [21, p. 2; see also 19]. Are serious games so different from previous instructional innovations or has the serious

games community just forgotten its own history? Importantly, in a different neighboring discipline, that of multimedia learning, the research outcomes, as organized by Mayer into the cognitive theory of multimedia learning (CTLM) [19], seem to be generally much clearer and unambiguous. Can DGBL studies be more like those underpinning CTLM than those about which Feinstein and Cannon complained?

One notable difference between these two types of studies is that the former tends to be conducted in a laboratory, while the latter takes place in a real-world setting. Especially digital games for the formal schooling system tend to be studied directly in schools and sometimes with real teachers. The reason for this is because the mere integration of a digital game into the formal schooling environment is difficult and its acceptance by the target audience is not guaranteed: many practical barriers to game integration exist; ranging from the unintelligibility of interfaces and game rules for some teachers and non-players, to a lack of access to equipment, e.g. up-to-date video cards, to barriers posed by fixed lesson times, e.g. [15, 17, 25]. Of course, the authentic context brings more confounding variables. Thus, it seems that two opposing things are needed at the same time: minimizing confounding variables; that is, moving studies to labs, while keeping the studies' external validity, i.e. running them in the real-world context. Can this tension be, at least to some extent, reconciled?

Since 2008 we have conducted five quasi-experimental DGBL studies in the context of secondary education; four of which were comparative and investigated learning effects (as opposed to mere acceptance of a game by students or teachers). These studies involved a total of more than 700 subjects. Based on these studies, and also based on general educational literature, we put together 6 Recommendations that we believe will help remove some of the confounding variables, yet allow for conducting a study in the real-world context or in a laboratory setting that closely approximates the authentic environment.

The goal of this paper is to present these 6 Recommendations. The paper should not be read as a definitive guideline for conducting DGBL studies. First, our advice or counsel refers to the technical aspects of running DGBL studies rather than conceptual issues related to formulating research questions and designing experiments. Second, the list is not exhaustive. Still, in our opinion, the Recommendations could bring us a step closer to the reconciliation of the laboratory/real-world tensions.

The paper proceeds as follows. Section 2 briefly introduces our studies. Section 3 outlines and describes our Recommendations. Section 4 presents a general discussion and our overall conclusions.

Fig. 1: Screenshots from StoryFactory, Orbis Pictus Bestialis, and Europe 2045.

2 Assumptions and Design of our Studies

These Recommendations can be applied to studies with experimental design similar to ours; namely to comparative studies of DGBL learning effects that combine various quantitative and qualitative measures. We now detail our studies' design.

The studies involved Europe 2045 [6], Orbis Pictus Bestialis (OPB) [5], Bird Breeder [22] and StoryFactory [4] (Fig. 1). Our team developed all of them except for Bird Breeder. Europe 2045 is a turn-based, complex, multi-player, strategy game for social science courses. It can be played during a one-day workshop or within a formal schooling framework on a long-term basis (about a month). OPB and Bird Breeder are simulation mini-games for explaining and practicing specific skills: animal training in the case of OPB; and Mendelian genetics in the case of Bird Breeder. Their goals can be achieved within 15 – 30 minutes of playing. Both games can be best used in the formal schooling system for homework assignments or in a practical seminar after a theoretical lecture on a specific topic (for gaining more insight into the topic). StoryFactory is not a game but a 3D toolkit, which helps students learn how to produce short movies in a 3D virtual world and is to be used in ICT/media education classes. The target audience for all the projects is students aged 13 and above.

Our first study (marked as EU1) investigated the acceptance of Europe 2045 as an educational tool by the target audience (N=220), without assessing real learning gains. Students' attitude towards the game was positive and the majority claimed that they learned more than or at least as much as they usually did [6]. Our other studies investigated learning gains, while comparing experimental groups taking part in DGBL activities to control groups receiving comparable "traditional" instruction. Knowledge of subjects was assessed using knowledge tests in (a) pre-test/series of post-tests or (b) immediate post-test/delayed post-test design. A study with OPB (N=100) (OPB1) showed no between-group difference right after the treatment but medium effect size positive learning gains for experimental groups in one month delayed post-tests [5]. However, our preliminary analysis of data from a consecutive study with OPB and Bird Breeder (OPB2) suggests that we failed to replicate these results (unpublished data; N=224). Similarly, our pilot study of the learning effects of Europe 2045 (EU2) showed mixed results [26] (N=153, note: some subjects participated in two studies featuring two different games). Some of our learning gain results, most notably from the OPB2 and EU2 studies, are difficult to interpret due to technical problems encountered while running the studies. Thus, the studies present excellent examples that help explain our Recommendations. The final study employed StoryFactory (SF) and the preliminary analysis suggests a small positive effect size of active exposure to StoryFactory compared to passive exposure (unpublished data). That was the only study conducted on high-school teachers (N=29).

In studies OPB1/2 and SF, after an initial theoretical lecture, each class was randomly divided into experimental and control groups. Both groups received two different treatments that were nevertheless comparable as relates to their educational content and time length. Study EU2 was longitudinal: a whole class of subjects played Europe 2045 about a month as part of their regular education. Therefore, random sampling was not possible. Instead game classes were matched with comparable control classes that also received a set of theoretical lectures on the European Union. In each study a teacher supervised DGBL activities.

3 Recommendations

This section summarizes our research and experimental experiences into Recommendations. In doing so, it verbalizes how to avoid several *technical* pitfalls during DGBL studies conducted in the context of the formal schooling system. Recommendations can be applied most straightforwardly to comparative studies in which classes of high-school students represent a pool of experimental subjects and every class receives a treatment at once or is divided into several groups, each of which receives a different treatment. The Recommendations do not discuss conceptual and methodological issues – how research questions should be formulated, what treatment should be picked for control groups, etc. The descriptions will follow this structure: summary – rationale – an example from one of our studies – take-home message.

Recommendation 1: Reserve a whole day for the experiment.

Summary. It is important to design the study so that subjects participate in research activities *only* during the school day on which the study takes place. In particular, (a) the treatment *per se* should start and end at approximately the same time for each class tested, and (b) students should not be involved in other educational activities not controlled by the researchers, nor should they be examined (no matter the subject), during the testing day. If possible, (c) subjects should not be involved in significant extra-curricular activities during the testing day or during the day prior to testing. Also (d) no major exams should take place the day after.

Rationale. Concerning Point (a), in general, subjects' overall mood and attention spans change during the day. Therefore, different performance levels can be expected if one class is examined early in the morning and another before lunch. Concerning Point (b), if students continue with their normal class work the day after the experiment, there is a high risk that some students' attention spans will be disrupted during the experiment (and therefore, performance will be compromised) due to the fact that they either worry they may be examined later that day or think about homework they need to finish for a class during that day. Concerning Point (c), significant extra-curricular activities, such as an official event the previous evening or one scheduled for the evening of that same day, may influence performance. The same applies for major exams the next day (d).

It can be argued that not respecting this Recommendation and conducting the study as part of a regular school day would actually increase external validity and – with a random sampling and a sufficiently large number of participants – would not pose problems for internal validity. While true, this argument is too idealistic. Because the possible influences are numerous and sometimes apply to individual participants, sometimes to part of a class and sometimes to a whole class. The number of classes should be an order of magnitude larger than it usually is in current DGBL studies, e.g. [1, 23, cf. 14], should the experiment be conducted in a "natural" setting. This is too costly. At the same time, our observation, which can be conceived as a working hypothesis, is that the more the experiment looks "laboratory-implemented," the more the high school students concentrate. We suspect that high school students are motivated *less* when undergoing a "natural" experiment (violating this Recommendation) than when undergoing a "laboratory" experiment or when doing the same tasks as part of their regular education. This Recommendation offers general advice on how to keep budgets reasonable at the cost of slightly reducing external validity.

Example. In the OPB1 study, which violated Points (a) and (b), some students tried to work on their homework during the experiment. Others did a sloppy job because they expected to be examined later that day and thus studied the subject in question instead of participating in the experiment (note: they might not have done their homework were this a regular lesson with OPB!). The OPB2 study violated Pt. (c): some students in one class left earlier because they participated in a competition that day.

Take home message. Fear of examination, the need to finish one's homework, fatigue from an important extra-curricular activity that occurred the previous day, or the fact that one is looking forward to an extra-curricular activity on the given day may distract students from experimental activities.

Recommendation 2. Disrupt the regular school schedule.

Summary: In many countries a formal schooling system implies a fixed schedule and lessons with a fixed duration. It is important to either accept this schedule completely, in particular to accept regular breaks, or disrupt this schedule entirely; ideally by taking students out of the school environment or by conducting a study on days when the regular schedule has been disrupted for the whole school.

Rationale. It may seem that if the experiment takes a whole school day (see Recommendation (1)), the schedule can be changed at the researchers' will. Unfortunately, this is not always the case. First, students often have their regular schedule *internalized*. Second, in some schools, the schedule is made explicit; e.g. by the bell ringing at the beginning and the end of lessons. Third, participants in the experiment have friends in different classes. If the schedule is not disrupted for these friends, they may come to visit the participants during periods that would normally be breaks. All of this means that if the experiment takes place in a real school environment and breaks in the experimental schedule are not matched with regular breaks, participant attention levels may decrease during the would-be regular breaks. It may also be undesirable to allow students from the control and experimental groups to mix together during breaks. Note that usage of a DGBL activity in school implies acceptance of the in-school schedule; however, researchers may want to change the schedule during the research day purely for experimental reasons, e.g. to introduce the experiment or to distribute questionnaires. In such cases, it can help to take students out of school and *model* the real breaks during DGBL activities (but not administer tests, etc.).

Example. In the OPB1 experiment, we wanted to test the effect of students interfacing with the OPB game after a regular expository lecture on the topic of that game. The expository lecture should have lasted the same time as it would have in the real educational setting; i.e. 45 minutes in the Czech Republic. However, before the expository lecture, we had to introduce the whole experiment to the class: that took 10 minutes. Thus, we had 55 minutes after which we wanted to schedule a break. However, as said, the regular lesson lasts 45 minutes. Oddly though, ten minutes before the end of the expository lecture, a high number of participants left to use the restroom and participants' friends started to wander into the class. Ultimately, we had to shorten the introduction to 5 minutes and the supplementary lesson to 40 minutes: a compromise.

Take home message. The regular school schedule should either be accepted or disrupted completely.

Recommendation 3. Work with a "standardized" expert teacher, who has authority over students.

Summary: Many DGBL activities should be supported by teachers, e.g. [12, 24]. According to our experience, the teacher effect has enormous influence on a study's outcome. Unless the teacher effect is investigated *per se*, all the classes and groups should have the same teacher; one who is, if possible, hypothesis-blind. The teacher should be an expert on the topic and should have authority over the students. Thus, inevitably, the teacher should be part of the research team.

Rationale: According to our experience, expertise levels among regular school teachers differ and teachers occasionally like/dislike particular topics. When supplementary activities are led/taught by different teachers, the outcome can be heavily influenced by the effect of the teacher's *a priori* knowledge, his/her authority, actual mood, attitude towards DGBL, etc.

Example: In the EU2 experiment, we let regular school teachers supervise and teach both the control and experimental groups. Some students (in the experimental group) later complained, in focus groups, that their regular school teacher did not engage less motivated students sufficiently so that they would participate in the game playing. This compromised the game play quality for the whole group (Europe 2045 is a multi-player game). At the same time, there was no similar problem with the control group. Eventually, in terms of learning effects, the data showed no difference between these two particular groups, but we got a positive gain for a different experimental group, compared to its matched control group supervised by a different teacher. It would be an important result, should it be proven that the game works for some teachers but not others. However, to investigate this hypothesis, enough classes must be available (i.e. around at least 30 group pairs) in order to gain statistically interesting results. If it is assumed that the teacher masters DGBL and just the effect of a game, or the presence of some of its features, are investigated, employing the same teacher during the trials would bring less noisy data.

Possible obstacles and solutions. The idea of all groups being led/taught by an expert teacher from the research team is not without drawbacks. However, according to our experience, such drawbacks are relatively minor compared to the teacher effect. First, the experimental design depicted on Fig. 2 (left) is impossible: the teacher cannot be in both groups at the same time. We suggest (and use) the experimental design on Fig. 2 (right). Both groups participate in the "introductory" lecture, but then one group is subject to a given intervention, while another waits. Then the groups are switched. For some group pairs the experimental group goes first. For others the control group goes first. Our suggestion for the "waiting" group is to use a supplementary activity that is not linked with the educational objective; i.e. one that is relatively easy but not boring, and is motivationally neutral or mildly positive. We used a five-factor personality test and informed the students that they would get the results, which mildly motivated most of them.

The second issue is, as revealed by our focus groups, that with an external teacher, some students may feel that they "do not need to learn as much as they would with their own teacher." As one student put it: "It was a pleasant change from the school routine. We didn't have to learn." (cf. Rec. 1). Generally, because our results showed some learning gains for most students in all our studies (though experimental groups did not always outperform their matched control groups), we do think that this issue is not as

troubling as the teacher effect. Arguably, the silent presence of the regular teacher during trials might help, but we do not have enough data to support this claim empirically.

The third point is that the teacher should be hypothesis-blind, which is not always possible if he/she is part of the research team. Two things can help reduce teacher bias to some extent: first, the teaching method should be "standardized". For instance, the teacher should have a set of key points that he/she should mention/focus on in both groups during the interventions. This can be checked by an in-class observer. Second, and more importantly, the research hypothesis should be formulated so that both negative and positive results are meaningful and the teacher's preferences for these outcomes are the same. Obviously, if the teacher is also the main author of the game, and the research question is merely whether this particular game outperforms a different educational activity, the bias is hardly avoidable.

Take home message. The "teacher effect" can corrupt the data. Unless the teacher effect is investigated, it is better to work with one expert teacher from the research team; i.e. one who is hypothesis-blind and has authority over students. If it is not possible for the teacher to be hypothesis-blind, every effort must be made to reduce possible bias.

Recommendation 4. Keep groups as small as possible for testing students.

Summary: If a DGBL activity is to take place in a real school with a whole class at once (but not e.g. as a homework), it is often desirable to set up the experiment so that the activity in the experiment also employs whole classes. This should be done no matter whether the game used is multi- or single-player. However, when tests are administered after the intervention, the groups should be broken into the smallest subgroups possible. These subgroups should not be allowed to interact with other subgroups when tested.

Rationale: Generally, the class effect tends to be large; similar to the teacher effect. In a field study, researchers are often interested in the class effect generated by the intervention; however, they are not interested in the class effect caused by *testing*. According to our experience, the latter is also large. When all (high school) students are tested at once, it is almost inevitable that some of them start joking during testing, saying right or wrong answers aloud, complaining and cheating (in our research, some students tried to cheat just for fun), or using mobile devices to find the right answers. So far, we have found no instruction that would help to ameliorate these problems. Note that some students always lack motivation to complete tests due to the low-stake problem (see Recommendation 5); however, the risk is high that these students would distract more motivated students and worsen *their* performance. Typically only a few students function as these "motivation reducers", which means that we cannot expect the performance reductions, caused by this effect, to cancel each other out in both paired groups. When students are tested in smaller subgroups, ideally groups of one person, the administrator can better control the subjects. The "motivation reducers" thus influence a smaller number of subjects (of course, an administrator must be present in each subgroup during the entire testing period).

Example: Our studies violated this Recommendation. For instance, in one OPB2 group (in an above-average high school according to the Czech School Inspectorate), one student started contemplating aloud about the administrator's sex life in hopes of attracting the attention of other students (which he eventually did).

Take home-message. It is better to administer tests and distribute questionnaires in subgroups smaller than the original research groups.

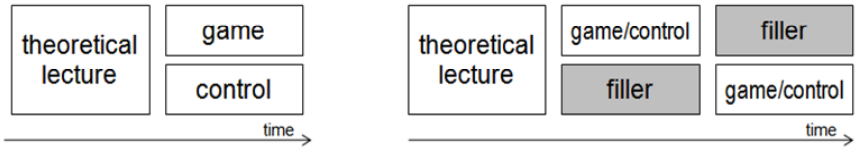

Fig. 2: Two possible designs, the one on the right features a "filling" activity.

Recommendation 5. Address the low-stake test problem.

Summary. The low-stake test problem is a general issue [28] and precautions should be taken to rectify it. Notably, (a) questionnaires/tests should not be too long and therefore should ask only the most important questions, (b) the order of questions should be changed in different versions of the tests, (c) the groups taking the tests in the same room should be as small as possible (Recommendation 4), (d) there must be no significant activity taking place after the test (Recommendation 1), (e) steps should be taken to avoid some students completing the test more quickly than others in the same subgroup.

Rationale. Generally, it is well known that in low-stake assessments that have little consequence for students, their motivation to complete the tests is reduced. This problem is discussed even in the context of large scale surveys such as the OECD Program for International Student Assessment (PISA), surveying student aptitude in reading, mathematics and science in more than 60 countries, e.g. [13]. Concerning (a) and (b), performance can decrease over time, for example, due to motivation or fatigue. Because of this, it is often useful, if a test is very long, to divide questions into blocks and present them in a different order in different versions of the test, so that the positions of question blocks are balanced across the test variants and research groups. This helps to separate the effect of fatigue and question difficulty.

Point (c) is actually part of Recommendation 4. Point (d) relates to Recommendation 1: student motivation is generally reduced if they expect a major event such as an exam. Special attention should be paid to delayed post-tests. It is tempting to think that delayed post-tests can be administered in one school class hour during a regular school day. Unfortunately, this is not the case. Ideally, the whole day should be reserved *only* for administering post-tests due to reasons mentioned in Recommendation 1. Because post-tests *per se* would hardly take a whole school day, it is necessary to supplement them with other activities unrelated to the regular school schedule.

Concerning (e), it often happens that some (often less motivated) students complete a test earlier than others. If the "earlier finishers" stay in the test room, they may distract the "late finishers." If the "early finishers" leave, the other students may speed up, thereby filling in tests less carefully. This is because they realize that they can leave as soon as they finish the test. Ideally, each student should have the same amount of time for every question, and each question should be put to all the students in one testing subgroup at the same time. This method is practiced by Mayer in his studies on learning from multimedia [19, p. 44]. However, according to our experience, it also helps when the whole test is divided into several sections, each of which is administered at the same time to every student in the testing group. If a student finishes a section, he/she waits only a few minutes for the next one (in Mayer's approach, every section consists of one question).

Finally, one should note that the official grading of the tests is not always the solution, since it is not possible to grade delayed post-tests (one has to administer them without notifying the students in advance; otherwise, they would study for the test). However, what might help is "gamification" of the testing process, i.e. embedding tests in a game, while making their filling an inherent part of the game-play.

Examples. In the OPB1 study, we measured students' knowledge with eight open-ended questions only. The test took about 15 minutes. For the OPB2 study, we reasoned that it would be advantageous to have twice as many open-ended questions to gain a more extensive sample of students' knowledge. The OPB2 test took about 30 minutes and the lessons learnt are that students left most questions unanswered or wrote just a few words instead of, as expected and required, several sentences. We should have used the shorter version and/or multiple-choice questions. At the same time, some students returned the test after 10 minutes hoping they could go to lunch, while the rest of the class continued working on the test. It was hard to keep the "early finishers" silent for the next 20 minutes. However, this became easier when the test was split into two parts; the second not being administered until 90% of the students had finished the first one (more parts would be even better).

In the SF study with high school teachers at a two-week-long summer school, 29 teachers completed immediate post-tests, but only seven volunteered to complete one week post-tests with eight questions, despite the fact that they could take home a bottle of wine. The test had four closed-ended and four open-ended questions and took 15 minutes. Notably, all teachers had many other things to do since the school year was ending.

Take-home message. The low-stake test problem can be partly addressed by giving questions to students one by one or in small batches and by avoiding administering the test when students are expecting a significant event after the test. It is also useful to change the order of the questions: this helps get balanced answers for all questions despite increasing fatigue/decreasing motivation. Gamification of tests may help too.

Recommendation 6. Avoid certain periods of year.

Summary. It is important to avoid conducting the experiment during certain periods of the school year, when low performance can generally be expected, e.g. before the end of the school year or during exam periods. The situation is complicated by the fact that it is often useful to administer delayed post-tests; these tests also should not be administered during the critical periods.

Rationale. The reason is the same as for Recommendations 1 and 5.

Example. We conducted the OPB2 and EU2 experiments in May 2011 and administered the delayed post-tests on roughly 25 June 2011. The school year ends on 30 June in the Czech Republic. While we administered the delayed post-tests, some students in some classes (but not in all classes) openly claimed that they would make no effort to complete the tests because their final marks had already been assigned (on about 20 June). Also, nearly half the students were missing in some classes. Students' average scores on one of the knowledge tests in the Europe 2045 experiment are depicted in Fig. 3. Because the tests were of the same difficulty and we also administered a 3-month delayed post-test (the school year starts on 1 September), attribution of the large decrease in the 1-month post-test scores for the experimental classes to the low-stake test issue seems valid. What's worse, preliminary analysis of the OPB2 study's data suggests that we in fact failed to replicate the OPB1 results (i.e. no difference in delayed post-tests). However,

because we do not have a 3-month post-test for the OPB2 study and the scores were really low for the 1-month post-test, we do not know whether to attribute the null results to a lack of difference between the instructional effectiveness of the DGBL activity and the control treatment or to the combination of the low-stake test problem and the floor effect.

Take-home message. It is important to consider the context in which the experiment and the knowledge assessment take place. Was the chosen time of year a distraction for the students?

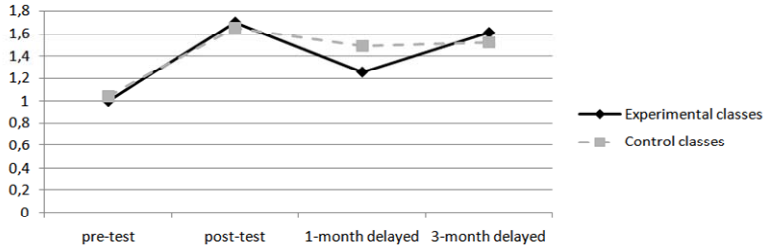

Fig. 3: Normalized averages across the experimental and control classes for the first type of knowledge test administered in the EU2 study (1 equals the average score across the control classes in the pre-test). Note: standard deviations are from 0.43 to 0.52 for all conditions and tests, except for the 1-month post-test, where they are larger: 0.57 (experimental) and 0.6 (control). Note also that the outcome of the second knowledge test is similar.

General social science recommendations. Besides the 6 Recommendations above specific to the DGBL, there are many standard social science recommendations. Newcomers to social sciences, such as computer scientists, should consult introductory text books, e.g. [3]. For instance, sometimes standardized knowledge tests are available. More often though researchers have to compile their own knowledge tests. In such cases, it is vital to identify and replace questions deemed "too easy" and "too difficult" as well as questions that do not distinguish able and less able students, see e.g. [16]. Another useful idea is to consider combining quantitative measures (such as questionnaires with Likert items and knowledge tests) with qualitative measures (such as commentary – e.g. on pictures or specific events) or text writing tasks for subsequent content analysis. While quantitative outcomes often provide a kind of ultimate aggregate description of *what* happened during the treatments, qualitative data can add interesting detail to this picture; they can help to elucidate *how* the gross quantitative outcome was achieved.

Even though these two (and other) general social science recommendations may seem obvious to social scientists, there are many DGBL studies, including our own OPB1 study [5], that do not mention how they (or that they) piloted the knowledge tests and many studies use only quantitative or qualitative methods.

4 Conclusion

When running a DGBL study on learning gains in the context of a formal schooling system, researchers often encounter many technical pitfalls stemming from the authenticity of the environment. In this paper, we have verbalized 6 Recommendations suggesting how to minimize some of these pitfalls:

1. Reserve a whole day for the experiment.
2. Disrupt the regular school schedule.
3. Employ a "standardized" expert teacher, who has authority.
4. Use the smallest groups possible for assessing students' knowledge.
5. Address the low-stake test problem.
6. Avoid certain periods of year.

Generally, Recommendations argue for running a study in a more laboratory-like fashion, while *modeling* the authenticity of the real school setting. This general suggestion can be interpreted as: turn the school into a laboratory, or model a school setting in a laboratory. Both ways help to eliminate some of the confounding variables.

The important boundary condition of the Recommendations is that they were formulated based on our experience with *secondary* education systems. Not all of these recommendations may apply in primary and tertiary education environments, e.g., the low-stake test issue may not be that problematic with younger children. However, some other issues may emerge, e.g., the necessity to keep the study short for primary school students.

We hope that the Recommendations will help to improve the quality of DGBL studies conducted in authentic environments, and that consequently, the empirical research base will become more solid.

Acknowledgments. This work was partially supported by the project *LEES: Learning Effects of Educational Simulations* nr. P407/12/P152 (GAČR) for C.B. and V.Š., and by the IGA MZČR grant NT/13386 for D.K. and D.L. We thank Zdeněk Hlávka, Tereza Nekovářová and Tereza Selmbacherová for their contributions to this research and three anonymous referees for their helpful comments.

References

1. Annetta, L. A., Minogue, J., Holmes, S. Y. and Cheng, M. Investigating the Impact of Video Games on High School Students' Engagement and Learning about Genetics, *Computers & Education*, 53(1), 74-85 (2009)
2. Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H.: Making Learning Fun: Quest Atlantis, A Game Without Guns. Educational Technology Research & Development 53(1) (2005) 86-107
3. Babbie, E. R.: The Practice of Social Research, 13th ed. Wadsworth Publishing (2012)
4. Bida, M., Brom, C., Popelova, M., Kadlec, R.: StoryFactory--A Tool for Scripting Machinimas in Unreal Engine 2 and UDK. In: *Proceedings of the ICIDS 2011*, LNCS 7069, Springer (2011) 334-337
5. Brom, C., Preuss, M., Klement, D.: Are Educational Computer Micro-Games Engaging And Effective For Knowledge Acquisition at High-Schools? A Quasi-Experimental Study. In *Computers & Education* 57 (2011) 1971-1988
6. Brom, C., Šisler, V. and Slavík, R.: Implementing Digital Game-Based Learning in Schools: the Augmented Learning Environment of Europe 2045. In: *Multimedia Systems*, 16(1) (2010) 23-41
7. Cuban, L.: Teachers and Machines: The Classroom Use of Technology Since 1920. New York: Teachers College Press (1986)

8. Egenfeldt-Nielsen, S.: Beyond Edutainment: Exploring the Educational Potential of Computer Games. PhD thesis. University of Copenhagen. (2005)
9. Feinstein, A. H., Cannon, H. M.: Construct of Simulation Evaluation. In: *Simulation & Gaming* 33(4) (2002) 425-440
10. de Freitas, S.: Learning in Immersive Worlds. Joint Information Systems Committee. (2006) Available: http://www.jisc.ac.uk/eli_outcomes.html [Accessed 16.3.2012].
11. Gee, J. P.: What Video Games Have to Teach Us About Learning and Literacy. New York: Palgrave/St. Martin's (2003)
12. Hays, R. T.: The Effectiveness of Instructional Games: A Literature Review and Discussion, *Technical Report 2005-004*, Orlando: Naval Air Warfare Center Training Systems Division (2005)
13. Hopfenberg, T. N.: Students' Test Motivation in PISA. In: *Proc. EARLI* (2011) 498-499
14. Huizenga, J., Admiraal, W., Akkerman, S. and ten Dam, G.: Mobile Game-based Learning in Secondary Education: Engagement, Motivation and Learning in a Mobile-city Game, *Journal of Computer Assisted Learning*, 25(4), 332-344 (2009)
15. Klopfer, E.: Augmented Learning: Research and Design of Mobile Educational Games. Cambridge: MIT Press (2008)
16. Izard J: Trial Testing and Item Analysis in Test Construction. Module 7. *Quantitative Research Methods in Educational Planning*. UNESCO International Institute for Educational Planning (2009)
17. Ketelhut, D. J., Schifter, C. C.: Teachers and Game-based Learning: Improving Understanding of How to Increase Efficacy of Adoption. In *Computers & Education* 56 (2011) 539-546
18. Malone, T. W.: Toward a Theory of Intrinsically Motivating Instruction. *Cognitive Science* 5(4) (1981) 333-369
19. Mayer, R. E.: Multimedia Learning. New York: Cambridge University Press (2001)
20. Mayer, R.E., Clark, R.C.: Simulations and Games in e-Learning. In: *E-Learning and the Science of Instruction*, 3rd. ed., Chap. 16, John Wiley & Sons (2011) 369-400.
21. Moreno, R.: *Instructional technology: Promise and pitfalls. In Technology-based education: Bringing researchers and practitioners together*, Greenwich, CT: Information Age Publishing. (2005) 1-19
22. Novak, M. and Wilensky, U.: NetLogo Bird Breeder Model. Available: http://ccl.northwestern.edu/netlogo/models/BirdBreeder. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (2007)
23. Papastergiou, M.: Digital Game-Based Learning in High School Computer Science Education: Impact on Educational Effectiveness and Student Motivation. In *Computers & Education*, 52, 1-12 (2009)
24. Sitzmann, T.: A Meta-analytic Examination of the Instructional Effectiveness of Computer-based Simulation Games. In *Personnel Psychology* 64 (2011) 489-528
25. Šisler, V., Brom, C.: Designing Educational Game: Case Study of Europe 2045. In *Transactions on Edutainment I*, Springer-Verlag Berlin Heidelberg (2008) 1-16
26. Šisler, V., Buchtová, M., Brom, C., Hlávka, Z.: Towards an Empirical–Theoretical Framework for Investigating the Learning Effects of Serious Games: A Pilot Study of Europe 2045. In *Applied Playfulness*. Proceedings of the Vienna Games Conference 2011: Future and Reality of Gaming. Braumüller Verlag. Vienna (2012) 16-36.
27. Wastiau, P. et al.: How are digital games used in schools? Complete Results of the Study. European Schoolnet. [Online]. (2009) Available at: http://games.eun.org/upload/gis-synthesis_report_en.pdf [Accessed: 16.3.2012].
28. Wise, S. L.: Strategies for Managing the Problem of Unmotivated Examinees in Low-Stakes Testing Programs. In: *The Journal of General Education* 58(3) (2009) 152-166
29. Wolfe, J., Crookall, D.: Developing a Scientific Knowledge of Simulation/Gaming. In: *Simulation & Gaming* 29(1) (1998) 7 -19