A Systematic Literature Review of Methodology Used to Measure Effectiveness in Digital Game-Based Learning

Anissa All, Elena Patricia Nuñez Castellar & Jan Van Looy
iMinds, MICT, Ghent University, Ghent, Belgium
Anissa.All@Ugent.be
ElenaPatricia.NunezCastellar@Ugent.be
J.Vanlooy@Ugent.be

Abstract: In recent years, a growing number of studies is being conducted into the effectiveness of digital game-based learning (DGBL). Despite this growing interest, however, it remains difficult to draw general conclusions due to the disparities in methods and reporting. Guidelines or a standardized procedure for conducting DGBL effectiveness research would allow to compare results across studies and provide well-founded and more generalizable evidence for the impact of DGBL. This study presents a first step in this process by mapping current practices through a systematic literature review.
The review included peer-reviewed journal and conference publications between 2000 and 2012. Other inclusion criteria were that (1) the study's primary aim was effectiveness measurement of cognitive learning outcomes, (2) the focus was on digital games and (3) a pre-post design with a control group was used. Twenty-five publications were found eligible for this study.
Important differences were found in the number of control groups used and the type of intervention implemented in the control group (e.g. traditional classroom teaching, use of multimedia, computer-based learning, paper exercises, other games, or no intervention). Regarding the implementation method of the DGBL intervention in the experimental group, two approaches can be distinguished: stand-alone intervention or as part of a larger program. Moreover, a wide variety of effectiveness measures was used: measures for learning outcomes were complemented with time measurements and/or with self-reported measurements for self-efficacy and motivation. Learning effect calculation also varied, introducing pre-test scores in the analysis, conducting a separate analysis on pre- and post-test scores or conducting an analysis on difference scores. Our study thus indicates that a variety of methods is being used in DGBL effectiveness research opening a discussion regarding the potential and requirements for future procedural guidelines.
Keywords: effectiveness, digital game-based learning, cognitive learning outcomes

1. Introduction

In recent years, attention for the use of digital games has grown in a wide range of sectors. Digital games that do not primarily aim at entertainment have been deployed in the field of education, health and wellbeing, government, NGOs, corporate, defence, marketing and communication (Sawyer and Smith 2008). This growing interest in digital game-based learning (DGBL) has resulted in an increasing amount of publications on the topic (Michael and Chen 2005). One important aspect in this field of research is effectiveness measurement (Connolly et al. 2012) whereby effects of DGBL and contributing elements on learning outcomes are assessed. An important limitation in this field is the incongruity of study designs (Kharrazi et al. 2012), which makes comparison across studies problematic. A consistent approach in effectiveness measurement would create the possibility to map important aspects of effectiveness on a more general level. Furthermore, uniformity in effectiveness studies on DGBL would help us gain better insight in validity and reliability of single studies. The present study takes a first step in the development of standardized guidelines by mapping the methods currently being used in effectiveness research on DGBL.

1.2. Defining effectiveness

In the literature, learning is often clarified on the basis of the generated outcomes (Gagne 1984). An effective instructional method can thus be described as a method which has a positive impact on learning achievement and therefore learning outcomes (Joy and Garcia 2000). An instructional method has been defined by Salomon (Cited by (Clark 1994) p. 23) as "Any way to shape information that activates, supplants or compensates for the cognitive processes necessary for achievement or motivation." Effectiveness of an instructional method can thus refer to either learning outcomes and/or motivation. According to Salomon (1993) the relationship between a medium used to teach and learning is an interaction between cognitive processes and characteristics of the mediatized environment. Medium and learning content are therefore inherently connected, implying that characteristics of the medium can influence the learning outcome (Kozma 1994).
A characteristic that has been detected as an important aspect in the learning potential of digital games is their intrinsically motivating character (Garris et al. 2002), meaning that the activity in itself is

engaging and no external reward for performing the activity is expected (Jenkins 2009). Intrinsically motivating activities create an enjoyable and fun experience, increasing the likelihood of repetitive usage (Ritterfeld et al. 2009).
Another aspect of effectiveness is transferability, which refers to the transfer of knowledge in a formal context to situations in real life (Kozma 1994). When the transfer between a formal context to real life situations is low, this is defined as inert knowledge (Whitehead 1959). According to several authors inert knowledge is often due to usage of traditional teaching methods, which are outdated in that respect (Renkl et al. 1996). Garris et al. (2002) state that, in the context of DGBL, this transfer can be stimulated by organizing a debriefing session after gameplay.

1.2. Effectiveness studies in DGBL

Typically, an experimental design is implemented to assess learning outcomes in a DGBL context by comparing a game-based approach with another type of instruction and/or no intervention. The types of interventions to which the game-based approach is compared can vary, which implies that results will ultimately depend on the particular comparison that is made (Bleumer et al. 2012). According to Campbell et al. (1963) the best experimental methodology for establishing whether learning has taken place is a pre-test post-test approach, including both an experimental and a control group.
Questionnaires are typically used to assess the motivational aspects of DGBL, gauging the motivations of participants for learning via the intervention received and their interest in participation (Hainey 2010). Questionnaires are also implemented to assess other affective outcomes, such as attitudes. Moreover, some studies use in-game assessment – referred to as stealth assessment – which is a technique that aims at accurately and dynamically measuring the player's progress (Shute et al. 2011). Finally, qualitative methods such as interviews and observation have also been used in the context of effectiveness studies of DGBL.
Three types of effectiveness studies in DGBL can be distinguished based on learning goals embedded in digital games (Bleumer et al. 2012). Specifically, digital games can aim at either knowledge transfer (cognitive learning outcomes), skill acquisition (skill-based learning outcomes) or attitudinal and behavioural change (affective learning outcomes). Games aimed at knowledge transfer are typically implemented in education. For example, some studies have found a positive impact of the use of digital games to teach math (Bai et al. 2012) and language (Yip and Kwan 2006). Digital games aimed at skill acquisition are typically implemented in a training and corporate context. Several studies have observed an impact of playing games to practice managerial skills (Corsi et al. 2006). Games aimed at behavioural change are typically implemented in the health sector. An example of this are the healthy eating games influencing the diet and physical activity of children (Baranowski et al. 2008). Games aimed at attitudinal or behavioural change are implemented to raise awareness on a certain topic, such as poverty (Neys et al. 2012).
According to Kraiger et al. (1993) these different types of learning outcomes require different types of assessment. Including studies aimed at the three learning outcomes would result in an extra level of heterogeneity, depending on the type of outcome that is assessed. Therefore, we will focus on one type of learning outcome in this study, that is cognitive learning outcomes.

2. Method

In the present study the Cochrane method was used to carry out our systematic literature review (Higgins et al. 2008). This review method has its origins in health research and aims to study the effectiveness of interventions for prevention, treatment and rehabilitation. According to Cochrane, four dimensions of study characteristics can be distinguished: 1) participants (e.g. characteristics of the sample involved), 2) intervention (contents, format, timings and treatment lengths, intervention(s) in control group(s)), 3) methods (e.g. applied research methods) and 4) outcome measures (e.g. instruments used to measure a certain outcome) and results (Higgins et al. 2008). This distinction was also made in the present study.

Search engines used for our review were Web of Knowledge, EBSCO Host and the International Bibliography of the Social Sciences. The following search string was used: ((Edu* OR serious OR learn* OR digital game based learning) AND ((dig* OR video OR computer) AND game) AND (assess* OR effect* OR measur*)). This search identified 54 publications dealing with effectiveness of DGBL aimed at cognitive learning outcomes. The review included peer reviewed journal and conference publications between 2000 and 2012. Other criteria for inclusion were that (1) the study's primary aim was effectiveness measurement of cognitive learning outcomes, (2) the focus was on digital games and (3) a pre-post design with a control group was used. Eight studies had a post-only design with a

control group and 21 studies had a pre-post design without a control group which were all excluded. Eventually, 25 studies with a pre-post design and control group were considered eligible for analysis.
A quantitative content analysis was conducted using SPSS. The codebook for this analysis was created inductively, using qualitative coding in nVivo. For this, open and axial coding (Glaser and Strauss 2009) were used for analysing procedure and methods sections of the studies based on Cochrane guidelines.

3. Results

3.1. Participants

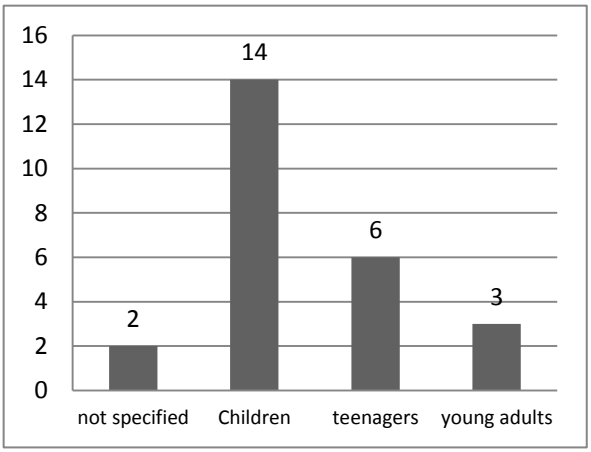

Figure 1: Subjects included in study (n = 25)

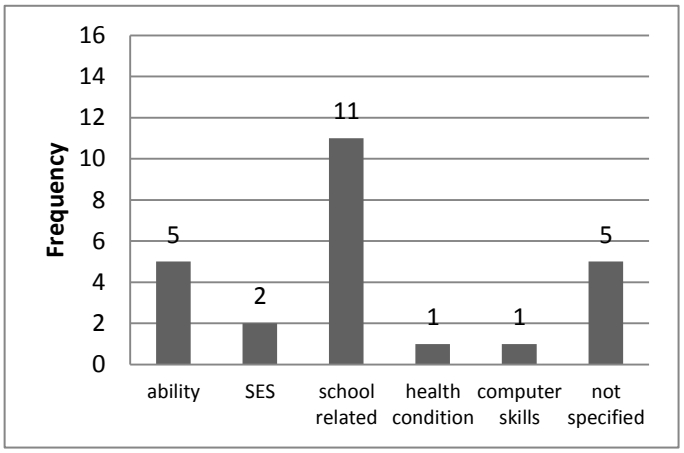

Figure. 2: Inclusion criteria (n = 25)

The average sample size of participants in studies reviewed was 220 (SD = 284). Although not all the studies reported the number of participants included by group (8% did not), our results showed that when reported the average number of participants was 105 (SD = 163) in the experimental and 84 (SD = 92) in the control group. Although four studies reported participants' mean age, most studies defined subjects based on types of people, such as 'university students'. Sixty-five per cent of the studies included children, 24% teenagers and 12% young adults (Figure 1). Inclusion criteria for participation were thus mostly school-related (e.g., 'majoring in math and science). Several studies only included a certain subgroup, including participants based on ability (e.g., low achievers), socioeconomic status or a certain health condition (Figure 2).

3.2. Intervention

Experimental groups (EG) were compared to a control group (CG) that either included participants that did not get an intervention (24%), got an intervention using another instructional approach (56%), or were compared to several control groups, combining both (16%). One study did not provide any information on interventions implemented in the CG (Table 1).

Table 1: Interventions in control group (n = 25)

Intervention in control group(s)	N	%
Traditional classroom teaching	12	48
Traditional classroom teaching, with the use of multimedia	1	4
Computer-based application, such as an educational website	4	16
Other game not related to the subject of the game implemented in the EG	2	8
Paper and pencil exercises	3	12
No intervention	10	40
Not specified	1	4

In the larger part of the studies (64%) DGBL was implemented in a formal context (e.g., in school during school hours), 8% in an informal context (e.g., home setting) and 12% in a semi-formal context (Figure 3) referring to an implementation in a formal institution, such as a school, but where gameplay occurred outside of school hours. Sixteen per cent did not specify the context of play and 56% did not specify the gameplay composition (Figure 4). Twenty-four per cent let participants play individually, 4% individually in competition, 24% cooperatively and 4% in a cooperative competition, meaning groups of

participants played together against other groups of participants. One study implemented all four gameplay conditions.

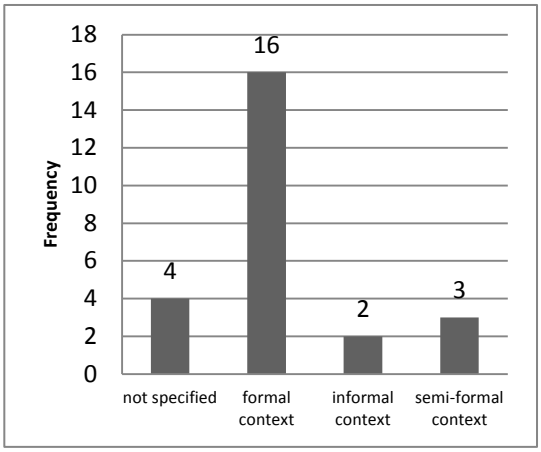

Fig. 3: Context of play (n = 25)

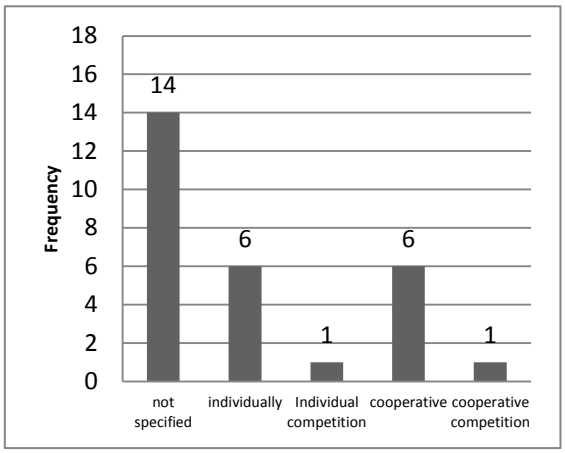

Fig. 4: Gameplay composition (n = 25)

Forty per cent of the studies did not report on the presence of an intermediary, referring to a teacher or researcher present during gameplay. In 56% of the studies an intermediary was present. One study did not include an intermediary. The average implementation period was 9 days (SD = 6), with a minimum of 1 day and a maximum of 23 days. Average total interaction time with the game is 12.4 hours (SD = 14.8), with a minimum of 30 minutes and a maximum of 64 hours.
Games were either implemented as a stand-alone intervention (28%) or were embedded in a larger program (48%). Twenty-four per cent did not specify implementation. Table 2 gives an overview of program specifications.

Table 2: Specifications about games embedded in a larger program (n = 12)

Program specifications	***N***	***%***	***Description***
Introduction	5	20	An introduction concerning game content and gameplay was provided by an intermediary. This does not refer to an in-game introduction
Training of participants before intervention	5	20	A training session before the intervention was provided
Extra material	8	32	Extra material such as articles, extra exercises, extra reading material, etc. were freely available
Online platform	3	12	The game was part of a larger educational online platform
Game task formulation	1	4	Certain tasks were formulated during gameplay
Required reading	2	5	The participants were expected to read next to gameplay
Procedural help by intermediary	3	12	The participants received help concerning the actual gameplay. This does not relate to content
Guidance by intermediary	3	12	The participants received guidance during gameplay in order to contextualize the game in the broader learning context
Supplement of course	6	24	Gameplay occurred next to the classes
Debriefing	3	12	A debriefing session was provided

Several studies implemented the game as a supplement of a course. However, half of these provided extra time for the experimental group to interact with the game in addition to the courses, therefore spending additional time with the learning content.

3.3. Method

All studies reviewed implemented an experimental design. Forty-for per cent used a randomized controlled trial; 24% randomly assigned subjects while 20% randomly assigned classrooms to one of

the conditions. Twelve per cent did not randomly assign participants to experimental and control group(s), but 'matched' participants in groups based on certain characteristics such as previous test scores, and 44% did not specify on group assignment of participants.

3.4. Measures

Less than half (44%) implemented standardized tests, six of these only used standardized tests while 5 studies combined standardized tests with tests developed by the researchers. Twenty per cent of the studies reviewed only implemented tests developed by the researchers and 24% used school tests or exams ('student achievement') as an accuracy measure. Two studies used both test scores and student achievement as an accuracy measure.

Twenty-eight per cent did not report on the similarity between the pre- and post-test measurements. Forty per cent employed the same test before and after the intervention, 8% changed the sequence of the questions and 8% used a similar test (e.g., other questions with the same type and difficulty levels). The latter did not report on how similarity of parallel tests were assessed. Sixteen per cent used a dissimilar pre- and post-test, such as midterm exam scores and final exam scores. Two studies also implemented a mid-test and four studies a follow-up test. Table 3 gives an overview of measures used in the studies.

Thirty-six per cent of the studies reported on how scoring on tests occurred. Three studies (12%) included an independent coder, of which two controlled for inter-rater reliability. One study used several, non-independent coders to control for inter-rater reliability.

Table 3: Measures used for determining effectiveness (n = 25)

Objective measurements	**N**	**%**
Accuracy	**19**	**76**
Test scores	16	64
Student achievement	5	24
Time measurements	**2**	**8**
Time on task	2	8
Subjective measurements	**N**	**%**
Self-measurements	**8**	**32**
Self-efficacy topic	4	16
Self-efficacy general	2	8
Perceived educational value	2	8
Motivation	**10**	**40**
Motivation towards educational intervention	7	28
- Post-only, EG	3	12
- Post-only, EG and CG	2	8
- Pre- and post, EG and CG	2	8
Motivation towards learning/educational content	3	12
- Post-only, EG and CG	2	8
- Pre-post, EG and CG	1	4
Other	2	8
Attitudes towards school	1	4
Teacher expectations	1	4

The larger part of the studies (76%) did a check on pre-existing differences between experimental and control group(s) and 36% of the studies included in this review reported on effect size. Twenty-four per cent did not report on statistical analysis. Table 4 shows how analysis of tests occurred.

Table 4: Data-analysis (n = 18)

Data analysis	N	%	Description	Example from studies reviewed
Absolute test scores comparison	7	28	Absolute pre-test and post-test scores of EG and CG are compared separately	*…the independent samples t-test was applied to examine whether the differences between the mean scores of the control and experimental groups in the pre-test and post-test were statistically significant* (Yip and Kwan 2006)
Absolute test scores, adding pre-test scores to the analysis	13	52	Absolute test scores of EG and CG are compared, taking the pre-test scores into account	*…pre-test scores on the specific subject tested were introduced as covariates in order to control for initial levels of the ability'* (Rosas et al. 2003)
Difference scores	**10**	**40**		
Item accuracy	1	4	Use of a specific scoring system	*Each factor was rated -1 if performance changed from correct to incorrect, 0 if there was no change and +1 if it changed from incorrect to correct. Then, each subject's scores were summed to create a summary difference score…*(Coles et al. 2007)
Gain/loss scores	5	20	The number of points the gained/lost between the pre-and post-test	*…paired-samples t tests were conducted to compare the treatment and control gain scores from pre-test to post-test…*(Kebritchi et al. 2010)
Percentage of improvement	4	16	Percentage of improvement between pre-and post-test	*…the percentage of improvement was calculated from the primary scores by subtracting the pre-test result from the post-test result and then dividing the difference by the maximum result of the test* (Ketamo 2003)
Error rates	**2**	**8**		
Out-of-game error rates	1	4	Number of mistakes made in the pre- and post-test are compared	*the educational effect…by comparing the number of mistakes of the students of the VR-ENGAGE sub-groups with the number of mistakes of the students of the respective sub-groups that had used the simple ITS* (Virvou et al. 2005)
In-game error rates	1	4	In game measurement of number of errors during gameplay are compared	*…the position and location of the mouse onscreen were recorded every 10th second. How successful children were at solving the computer assignments immediately or after one or more repetitions can be derived from these registrations* (Van Der Kooy-Hofland et al. 2012)

4. Discussion

The results of the present study show that studies vary on different dimensions of the study design, presenting a heterogeneity of methodologies. Homogeneity is, however, an important prerequisite when conducting meta-analyses, impeding generalizing conclusions on the effectiveness of DGBL (Higgins et al. 2008). Differences were not only found between study designs, but also in reporting.
Regarding the participants dimension, several studies only used certain subgroups (e.g. certain ability, certain socioeconomic status). This does logically narrow results to this specific subgroup (Campbell et al. 1963), which is problematic, however, when generalizing claims on DGBL effectiveness are made. Therefore, reporting on inclusion criteria for recruitment and how sampling occurred, is essential.

When considering the intervention dimension, studies firstly differed on the type of intervention implemented in the CG. The interpretation of the contribution of the intervention to the EG does, however, depend on the activities performed in the CG (Campbell et al. 1963). Considering that intervention in the CG can influence results and interventions implemented in CG differed across studies, comparison between results becomes problematic. Secondly, implementation of DGBL in the EG differed between studies, either implementing them as a stand-alone intervention or in a larger program. When embedded in a program, elements of the program differed across studies as well (e.g. introduction, debriefing, extra material, required reading, etc.). The addition of other elements to the intervention can result in multiple treatment interference (Campbell et al. 1963), however, meaning the achievement gain might not be solely attributable to DGBL, but could be influenced by other activities that are part of the intervention. Results of studies implementing only DGBL and studies implementing DGBL in a larger program are thus not comparable. Thirdly, implementation of DGBL differed by the presence of an intermediary and the role of this intermediary. Most studies did not report on whether or not an intermediary was present. When an intermediary was present, they were either present to

solely supervise or were present with the purpose of providing procedural help and/or guidance during gameplay. The role of the intermediary was either filled by the classes' teachers or a researcher. Who the intermediary is (e.g. someone more familiar such as a teacher or a total stranger) and how he or she interacts with the participants is a potential confound when assessing the effect of the DGBL intervention (Leary 1995).

While all studies implemented an experimental design, differences were found in the participants' assignment to the EG and CG which was done with or without randomization or by 'matching' in order to attain similarity of both groups. It was, however, not clear whether this matching occurred randomly. When using matched random assignment, the participants' scores on a measure of relevance (for example: pre-test) are obtained in order to randomly assign participants belonging to a certain level to the conditions (Leary 1995). According to Campbell & Stanley (1963), matching is not a preferable method. In the context of educational research, randomization of the classroom as a unit is more preferable, because classrooms can then be classified for analysis on the basis of factors such as schools, teacher, subject, time of day, mean intelligence level, etc. (Campbell et al. 1963). According to Leary (1995), however, randomization of schools will jeopardize internal validity, as groups will likely differ on multiple dimensions. This is an issue that merits further discussion, considering randomized controlled trials are difficult to implement with small sample sizes, which are often a reality in DGBL effectiveness studies.

How effectiveness was measured, also differed between studies, complementing learning outcome measures with affective measures and time measures. Some studies reported very specifically on how scoring on tests occurred in. According to Campell & Stanley (1963) different instruments and different scorers, can yield other results. Every measuring tool also entails a certain measurement error, fuelled by transient states (e.g. participants' mood, level of fatigue), stable attributes (e.g. misunderstanding questions, individual differences in motivation), situational factors and characteristics of the measurement itself (e.g. ambiguous questions, test that induce fatigue). Incomplete information on and differences between the length of certain tests, formulating of questions or the time of measurement account for an incomparability of results across studies as well. Use of independently developed standardized tests could provide more 'stable' measurements and create comparability across studies. This is, however, a difficult exercise, considering the wide range of topics covered by DGBL. Although, standardized tests could be implemented to assess affective learning outcomes, such as motivation.
While the larger part of the studies implemented the exact same test pre- and post-intervention, others changed the sequence of the questions. Implementing the same questions in the pre-and post-test can however lead to a test-retest practice effect (Campbell et al. 1963). According to Crawford et al. (1989) this is due to retention of specific test material by the participants. Other studies used similar tests, meaning these consisted of questions of the same type and difficulty level. While practice effects can still occur using a parallel version of a test on different points in time (e.g. pre- and post-test), these generally tend to be smaller (Anastasi 1961). Certain studies also used dissimilar tests, when for example student achievement in school (e.g. exam scores) was used as a measure. This seems problematic, considering assumptions on the comparability of both tests cannot be made, making any significant achievement gains possibly invalid. Differences in similarity of pre-and post-test across studies, is another reason why it is difficult to compare results across studies.

Important differences were also found when considering data analysis techniques. While indeed several analyses can be used to measure the effect of the intervention, an analysis of covariance (ANCOVA) with pre-test scores as a covariate, is a more preferable method (Campbell et al. 1963). In the context of randomized controlled trials, ANCOVA reduces error variance and in the context of nonrandomized designs, it adjusts mean scores of the post-test to differences between groups on pre-test scores (Dimitrov and Rumrill 2003).
Furthermore, missing information on implementation of the intervention(s) impedes replication of certain studies, which is a basic principle of empirical research in order create the opportunity to falsify obtained results (Popper 2000). Finally, incomplete information on sampling, similarity between interventions of the EG and CG and similarity between pre-and post-tests, puts validity of certain results in doubt.

In general, we can conclude that comparisons between studies are problematic as a result of heterogeneity in study designs, heterogeneity in reporting, incomplete information and biased results, impeding generalization. Standardized guidelines on sampling, activities in control groups,

implementation of DGBL in the experimental group, measures, scoring, analysis and reporting on these elements could contribute to homogeneity in the research field and create insight in the validity of studies.

5. Limitations and further research

The selection and coding of publications was conducted by one researcher, which can be considered a limitation of this study. This study also is limited to digital games aimed at cognitive learning outcomes. Further research should thus be conducted on methodologies used in digital games aimed at skill acquisition and behavioural or attitudinal change.

An interesting venue for future research is exploring the possibilities for the development of a standardized procedure to measure effectiveness of DGBL. Relevant issues to investigate in this context are gathering input from experts in the methodology field in order to detect preferable methods for measuring learning effectiveness (e.g. number of control groups, activity in control group, implementation of DGBL, implementation period, etc.). Further, such a procedure should be adjusted to the requirements of the people who would benefit from this procedure and actually use the procedure. Therefore, involvement of relevant stakeholders in the process of developing the procedure is desirable.

6. Acknowledgements

This PhD project is funded by IWT, the Flemish government agency for Innovation by Science and Technology (IWT).

7. References

Anastasi, A., 1961. Differential psychology: Individual and group differences in behavior. Macmillan.

Bai, H., et al. 2012. Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students. British Journal of Educational Technology, 43(6), 993-1003.

Baranowski, T., et al. 2008. Playing for real: video games and stories for health-related behavior change. American journal of preventive medicine, 34(1), 74.

Bleumer, L., et al., 2012. State of play of digital games for empowerment and inclusion: a review of the literature and empirical cases. Spain.

Campbell, D. T., Stanley, J. C. and Gage, N. L., 1963. Experimental and quasi-experimental designs for research. Houghton Mifflin Boston.

Clark, R. E. 1994. Media will never influence learning. Educational Technology Research and Development, 42(2), 21-29.

Connolly, T. M., et al. 2012. A systematic literature review of empirical evidence on computer games and serious games. Computers & Education.

Corsi, T. M., et al. 2006. The real-time global supply chain game: New educational tool for developing supply chain management professionals. Transportation Journal, 61-73.

Crawford, J., Stewart, L. and Moore, J. 1989. Demonstration of savings on the AVLT and development of a parallel form. Journal of Clinical and Experimental Neuropsychology, 11(6), 975-981.

Dimitrov, D. M. and Rumrill, J., Phillip D 2003. Pretest-posttest designs and measurement of change. Work: A Journal of Prevention, Assessment and Rehabilitation, 20(2), 159-165.

Gagne, R. M. 1984. Learning outcomes and their effects: Useful categories of human performance. American Psychologist, 39(4), 377.

Garris, R., Ahlers, R. and Driskell, J. E. 2002. Games, Motivation, and Learning: A Research and Practice Model. Simulation & Gaming, 33(4), 441-467.

Glaser, B. G. and Strauss, A. L., 2009. The discovery of grounded theory: Strategies for qualitative research. Transaction Books.

Hainey, T., 2010. Using Games-Based Learning to Teach Requirements Collection and Analysis at Tertiary Education Level.

Higgins, J. P., Green, S. and Collaboration, C., 2008. Cochrane handbook for systematic reviews of interventions. Wiley Online Library.

Jenkins, H., 2009. Confronting the challenges of participatory culture: Media education for the 21st century. The MIT Press.

Joy, E. H. and Garcia, F. E. 2000. Measuring Learning Effectiveness: A New Look at No-Significant-Difference Findings. JALN, 4(1), 33-39.

Kharrazi, H., et al. 2012. A Scoping Review of Health Game Research: Past, Present, and Future. Games for Health Journal, 1(2), 153-164.

Kozma, R. B. 1994. Will Media Influence Learning? Reframing the Debate. . Educational Technology Research and Development, 42(2), 7-19.
Leary, M. R., 1995. Introduction to behavioral research methods. Brooks/Cole Pacific Grove, CA.
Michael, D. R. and Chen, S. L., 2005. Serious games: Games that educate, train, and inform. Muska & Lipman/Premier-Trade.
Neys, J., et al., Poverty is not a game: behavioral changes and long term effects after playing PING. ed. 13th annual conference on the International Speech Communication Association, 2012 Portland.
Popper, K. 2000. Science: conjectures and refutations. Readings in the Philosophy of Science: From Positivism to Postmodernism, 9-13.
Renkl, A., Mandl, H. and Gruber, H. 1996. Inert knowledge: Analyses and remedies. Educational Psychologist, 31(2), 115-121.
Ritterfeld, U., Cody, M. and Vorderer, P., 2009. Serious games: mechanisms and effects. New York: Routledge.
Salomon, G. 1979. No distribution without individuals' cognition: a dynamic interactional view.
Sawyer, B. and Smith, P. 2008. Taxonomy for Serious Games. Digitalmil, Inc& Serious Games Initiative/Univ. of Central Florida, RETRO Lab.
Shute, V. J., Rieber, L. and Van Eck, R. 2011. Games... and... learning. Trends and issues in instructional design and technology, 3.
Whitehead, A. N. 1959. The aims of education. Daedalus, 88(1), 192-205.
Yip, F. W. M. and Kwan, A. C. M. 2006. Online vocabulary games as a tool for teaching and learning English vocabulary. Educational Media International, 43(3), 233-249.

8. Appendix: Studies included in review

Anderson, J. and Barnett, M. 2010. Using Video Games to Support Pre-Service Elementary Teachers Learning of Basic Physics Principles. Journal of Science Education and Technology, 20(4), 347-362.
Bai, H., et al. 2012. Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students. British Journal of Educational *Technology,* 43(6), 993-1003.
Coles, C. D., et al. 2007. Games that "work": using computer games to teach alcohol-affected children about fire and street safety. Res Dev Disabil, 28(5), 518-530.
Din, F. S. and calao, J. 2001. The effects of playing educational video games in kindergarten achievement. . Child Study Journal, 31(2), 95-102.
Kajamies, A., Vauras, M. and Kinnunen, R. 2010. Instructing Low-Achievers in Mathematical Word Problem Solving. Scandinavian Journal of Educational Research, 54(4), 335-355.
Kanthan, R. and Senger, J.-L. 2011. The Impact of Specially Designed Digital Games-Based Learning in Undergraduate Pathology and Medical Education. The Impact of Specially Designed Digital Games-Based Learning in Undergraduate Pathology and Medical Education, 135, 135-142.
Ke, F. 2008. Computer games application within alternative classroom goal structures: cognitive, metacognitive, and affective evaluation. Educational Technology Research and Development, 56(5-6), 539-556.
Kebritchi, M., Hirumi, A. and Bai, H. 2010. The effects of modern mathematics computer games on mathematics achievement and class motivation. Computers & Education, 55(2), 427-443.
Ketamo, H. 2003. An Adaptive Geometry Game for Handheld Devices. Educational Technology & Society, 6(1), 83-94.
Lorant-Royer, S., et al. 2010. Kawashima vs “Super Mario”! Should a game be serious in order to stimulate cognitive aptitudes? Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 60(4), 221-232.
Miller, D. J. and Robertson, D. P. 2010. Using a games console in the primary classroom: Effects of ‘Brain Training’ programme on computation and self-esteem. British Journal of Educational Technology, 41(2), 242-255.
Miller, D. J. and Robertson, D. P. 2011. Educational benefits of using game consoles in a primary classroom: A randomised controlled trial. British Journal of Educational Technology, 42(5), 850-864.
Moreno, J. 2012. Digital Competition Game to Improve Programming Skills. Educational Technology & Society, 15(3), 288-297.
Moshirnia, A. 2007. The Educational Potential of Modified Video Games. Issues in Informing Science and Information Technology, 4, 511-521.
Papastergiou, M. 2009. Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation. Computers & Education, 52(1), 1-12.
Parchman, S. W., et al. 2000. An Evaluation of Three Computer-Based Instructional Strategies in Basic Electricity and Electronics Training. . Military Psychology, 12(1), 73-87.

Poli, D., et al. 2012. Bringing Evolution to a Technological Generation: A Case Study with the Video Game SPORE. The American Biology Teacher, 74(2), 100-103.
Rastegarpour, H. and Marashi, P. 2012. The effect of card games and computer games on learning of chemistry concepts. Procedia - Social and Behavioral Sciences, 31, 597-601.
Rosas, R., et al. 2003. Beyond Nintendo: design and assessment of educational video games for first and second grade students. Computers & Education, 40, 71-94.
St Clair-Thompson, H., et al. 2010. Improving children's working memory and classroom performance. Educational Psychology, 30(2), 203-219.
Suh, S., Kim, S. W. and Kim, N. J. 2010. Effectiveness of MMORPG-based instruction in elementary English education in Korea. Journal of Computer Assisted Learning, 26(5), 370-378.
Van der Kooy-Hofland, V. A., Bus, A. G. and Roskos, K. 2012. Effects of a brief but intensive remedial computer intervention in a sub-sample of kindergartners with early literacy delays. Read Writ, 25(7), 1479-1497.
Virvou, M., Katsionis, G. and Manos, K. 2005. Combining Software Games with Education: Evaluation of its Educational Effectiveness. Educational Technology & Society, 8(2), 54-65.
Yang, Y.-T. C. 2012. Building virtual cities, inspiring intelligent citizens: Digital games for developing students' problem solving and learning motivation. Computers & Education, 59(2), 365-377.
Yip, F. W. M. and Kwan, A. C. M. 2006. Online vocabulary games as a tool for teaching and learning English vocabulary. Educational Media International, 43(3), 233-249.