Measuring Effectiveness in Digital Game-Based Learning: A Methodological Review

Article *in* International Journal of Serious Games · May 2014

DOI: 10.17083/ijsg.v1i2.18

CITATIONS
53

READS
3,780

4 authors, including:

Anissa All
iMinds-MICT-Ghent University
31 PUBLICATIONS **526** CITATIONS
SEE PROFILE

Elena Patricia Nunez Castellar
Ghent University
37 PUBLICATIONS **676** CITATIONS
SEE PROFILE

Jan Van Looy
Ghent University
175 PUBLICATIONS **2,174** CITATIONS
SEE PROFILE

Measuring Effectiveness in Digital Game-Based Learning: A Methodological Review

[1]Anissa All, [2]Elena Patricia Nuñez Castellar, [3]Jan Van Looy
[*1,]*iMinds-MICT-Ghent University, Anissa.All@Ugent.be*
[2,] *iMinds-MICT-Ghent University, Elenapatricia.Nunezcastellar@Ugent.be*
[3,] *iMinds-MICT-Ghent University, J.Vanlooy@Ugent.be*

Abstract

In recent years, a growing number of studies are being conducted into the effectiveness of digital game-based learning (DGBL). Despite this growing interest, there is a lack of sound empirical evidence on the effectiveness of DGBL due to different outcome measures for assessing effectiveness, varying methods of data collection and inconclusive or difficult to interpret results. This has resulted in a need for an overarching methodology for assessing the effectiveness of DGBL. The present study took a first step in this direction by mapping current methods used for assessing the effectiveness of DGBL. Results showed that currently, comparison of results across studies and thus looking at effectiveness of DGBL on a more general level is problematic due to diversity in and suboptimal study designs. Variety in study design relates to three issues, namely different activities that are implemented in the control groups, different measures for assessing the effectiveness of DGBL and the use of different statistical techniques for analyzing learning outcomes. Suboptimal study designs are the result of variables confounding study results. Possible confounds that were brought forward in this review are elements that are added to the game as part of the educational intervention (e.g., required reading, debriefing session), instructor influences and practice effects when using the same test pre- and post-intervention. Lastly, incomplete information on the study design impedes replication of studies and thus falsification of study results.

Keywords: *Digital Game-Based Learning, Effectiveness assessment, cognitive learning outcomes*

1. Introduction

In recent years, there has been a growing interest in the potential of games as instructional tools in areas such as education, health and wellbeing, government, NGOs, corporate, defense, marketing and communication [1]. Considering that the development and implementation of digital game-based learning (DGBL) implies a substantial financial effort, there is an increasing need to determine the educational potential of DGBL in order to justify the investment [2, 3]. One major justification of this investment should be well-founded empirical evidence [2]. While in recent years, there has been an increasing number of publications aimed at assessing the effectiveness of DGBL, there is still a lack of sound empirical evidence [4]. The lack of an overarching methodology for effectiveness research on DGBL has led to the use of different outcome measures for assessing effectiveness [5], varying methods of data collection [6] and inconclusive or difficult to interpret results [7]. Moreover, questions have been raised regarding the validity of current effectiveness research on DGBL [2, 5, 8]. A common methodology for assessing the effectiveness of DGBL would firstly create the opportunity to compare results and thus the quality of the different educational interventions across studies. Secondly, claims regarding the effectiveness of DGBL could be made on a more general level. Lastly, a common methodology could set a baseline for quality, which could serve as an evaluation tool for published studies and as a starting point for researchers desiring to conduct an effectiveness study on DGBL. The present study aims at mapping current research methods used for effectiveness research on DGBL and is a first part of a larger project aimed at the development of a standardized procedure for assessing the effectiveness of DGBL.

1.1 Defining effectiveness of DGBL

Based on the projected primary learning outcomes, three types of DGBL can be distinguished aiming at knowledge transfer (cognitive learning outcomes), skill acquisition (skill-based learning outcomes) or attitudinal/ behavioral change (affective learning outcomes) [9]. Games that primarily aim at knowledge transfer are typically implemented in education, in order to teach math [10] or language [11] for example. Digital games that primarily aim at skill acquisition are used for training, for example in a corporate or military context. Several studies have for instance examined the impact of playing games to practice managerial skills [12, 13]. Games aimed at attitudinal change are also used by governments and NGOs to raise awareness of a certain topic such as poverty [14]. Games aimed at behavioral change are typically found in the health sector, for example games promoting healthier food and physical activity to children [15]. Learning is, however, a multidimensional construct and while DGBL can primarily aim at a certain type of learning outcome, it can entail secondary learning outcomes. For instance, a game that primarily aims at teaching children English (cognitive learning outcomes) can also result in a more positive attitude towards learning English or English as a subject (affective learning outcomes).

According to O'Neill et al. [5] effectiveness of DGBL can be defined in terms of 1) intensity and longevity of engagement with a game 2) commercial success of a game and 3) acquisition of knowledge and skills as a result of the implementation of a game as an instructional medium. In the current study, we will focus on the third aspect, and more specifically on the acquisition of knowledge.

The effectiveness of DGBL as an instructional medium firstly consists of first order learning effects, referring to a direct influence on knowledge, skills, attitudes or behavior. This is typically assessed by looking at changes between pre- and post-game measurements [3]. A second aspect of effectiveness of DGBL is transfer, referring to the application of the learning content to real world situations [16]. This is typically assessed gathering data in the field, such as key performance indicators or by organizing a follow-up test [3]. As mentioned before, primary learning outcomes can entail certain secondary learning outcomes (e.g., a game that aims at teaching math skills can also lead to a more positive attitude towards math). In the case of educational interventions, especially when choosing for DGBL, motivation is often a secondary learning outcome one wishes to attain. Motivation is a necessary prerequisite to ensure that learners actually learn something. When they are not motivated, the chance of failing of an educational program will increase [17]. Moreover, according to Kozma [18], medium and learning content are inherently connected, implying that characteristics of the medium can influence the learning outcome. The power of games to intrinsically motivate players to engage in the activity (i.e., performing the activity in itself and for itself [19]) has been considered as an important aspect of games which can benefit learning [20]. More specifically, intrinsic motivation for performing an activity is associated with higher levels of enjoyment, interest, performance, higher quality of learning and a heightened self-esteem [21]. This type of motivation, however, is often assumed in the context of gaming, but is not always a reality. Especially in the context of DGBL, players can be extrinsically motivated to participate, referring to engaging in the activity as a result of external coercion. However, extrinsic motivation can be nuanced and subdivided in different types, depending on the extent to which their regulation is autonomous. The least autonomous form of extrinsic motivation is external regulation, which refers to an activity that is performed in order to receive a reward or avoid some negative contingency. An example of external regulation is engaging in DGBL in order to receive extra credits for a certain class. A more autonomous form of extrinsic motivation is introjected regulation and refers to an activity that is performed out of a sense of guilt or obligation or a need to prove something. Engaging in DGBL in a classroom context out of fear of negatively being evaluated by the teacher is an example of introjected regulation. The second most autonomous form of extrinsic motivation is identified regulation which refers to the performing of an activity, because the action or the outcome is accepted as personally important. An example of this type of regulation is engaging in DGBL for programming, because it will help the player to achieve his goal of becoming a programmer. Integrated regulation is the most autonomous type of external motivation and refers to regulations that are fully assimilated to the self and are consistent with other goals and values. For instance, when a pupil engages in DGBL in a school context, because he/she wants to be a good student, is an example of integrated regulation. These different types of extrinsic motivation are also associated with different outcomes and experiences. More specifically, higher levels of autonomy of extrinsic motivation result in higher levels of engagement, performance, higher quality of learning and lower levels of dropout. How autonomous the external motivation is, depends on the level of internalization of regulations or

values. Internalization of regulation and values can, however, be stimulated by the feeling of relatedness with significant others modeling or valuing a certain behavior. Perceived competence (i.e. self-efficacy) and the experience of autonomy (i.e. feeling of volition) [21] also play an important role in this internalization process.

2. Evaluation of educational interventions

Educational evaluation aims at describing and explaining experiences of students and teachers and judging the effectiveness of education [22]. Two types of evaluation can be distinguished: formative and summative evaluation. Formative evaluation aims at detecting areas for improvement, thus evaluating the process, whereas summative evaluation aims at determining to what extent an educational intervention was successful, thus judging its effectiveness [23]. While summative evaluation can occur independently, formative evaluation cannot occur without a summative evaluation [24].

Educational evaluation is not the same as educational research which requires more rigorous standards of reliability and validity [25]. Educational research can be conducted in two ways: by using a naturalistic design, describing an ongoing process in its natural setting, mostly by using observations or by using an experimental design which evaluates the impact of an educational intervention on its desired learning outcomes. DGBL effectiveness research should thus strive for more rigorous standards of validity and reliability in order to be considered as educational research, which underlines the need for defining standards.

3. DGBL effectiveness studies

The most implemented designs in DGBL effectiveness studies are quasi-experimental and survey design. A study of Chen and O'Neill [5] has shown that in most empirical studies on DGBL effectiveness, no pre-test of knowledge is implemented. According to Clark [2] the absence of a pre-test of knowledge is problematic, because differences in learning outcomes could be due to knowledge differences between individuals or groups at the start of the intervention. Consequently, this can lead to an overestimation of the instructional effect.

Moreover, when control groups are included in the studies, often no educational activity is implemented in the control group [5, 8]. According to Hays [8] the comparison to a control group, which does not receive an intervention or does not engage in educational exercises, is problematic in this type of research because, again, it might lead to an overestimation of the beneficial effects of DGBL. This is also supported by Clark [2] who states that one of the major motivations for the use of DGBL should be the justification of the investment made and should thus be compared to viable and less expensive alternative ways to teach the same knowledge and skills. According to Clark [2], this comparison should also be made on motivational aspects, and more specifically on motivation to learn through the game-based approach compared to other instructional programs.

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 [26]. Questions have been raised by several authors in the field about the validity of these measures [27] considering student opinion on for example learning and motivation has previously been found to be unreliable and conflicting with direct measures [2]. Suggestions have been made towards physiological or behavioral measures (e.g., eye-tracking, skin conductance), because data can be collected during game play in a more controlled manner [27]. Furthermore, motivation as a construct in the context of DGBL effectiveness research needs to be further examined since questions can be raised on whether definitions of motivation in different studies truly represent motivation or other constructs [27]. Further, questionnaires are also implemented to assess other affective outcomes, such as attitudes [27].

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, analyzing the player's competencies at various levels during game play [28]. Using technology, which strategies the player uses to solve certain problems can for instance be assessed in the game, giving the researcher information on the learner's progress [29]. Finally, qualitative methods such as interviews (e.g., attitudes before game play, player experiences after game play) and observation (e.g. behavioural performance after playing game, decision making and emotional reactions during game play) have also been used in the context of effectiveness studies of DGBL [3].

4. Method

In the present study the Cochrane method was used to carry out our systematic literature review [30]. 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, for dimensions of study characteristics can be distinguished: 1) participants (e.g., characteristics of the sample involved), 2) intervention (e.g., 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.

For this review, we only included studies that implemented games which primarily aim at cognitive learning outcomes, considering the different types of learning outcomes require different types of assessment and thus resist categorization in one research taxonomy [31].

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. Criteria for inclusion were that (1) the publications were peer-reviewed journal and conference publications between 2000 and 2012 (2) the focus was on digital games and (2) a pre-post design with a control group was used. According to Campbell et al. [32], a pre-post control group design is the best design to assess learning considering that a pre-test offers the opportunity to measure progress and a control group ensures us that this progress is not due to a mere lapse of time. 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 by coding the methods and procedures sections in the studies both deductively (fixed dimensions of study design based on Cochrane) and inductively (methods and elements belonging to dimensions of the study design) in nVivo. Open coding was used for identifying different methods and creating labels (e.g., randomization of subjects, randomization of classrooms, matching of participants). Subsequently, axial coding was used for creating categories by relating labels to each other representing different elements of the study design (e.g., assignment of participants). Lastly, the categories were assigned to the different dimensions of the study design as defined by Cochrane.

5. Results

5.1. Participants

Inclusion criteria for participation in the studies were mostly school-related (e.g., 'majoring in math and science). Other studies included a certain subgroup, including participants based on ability (e.g., low achievers), socioeconomic status (SES) or a certain health condition. Twenty per cent did not specify inclusion criteria used for participants (Fig. 1).

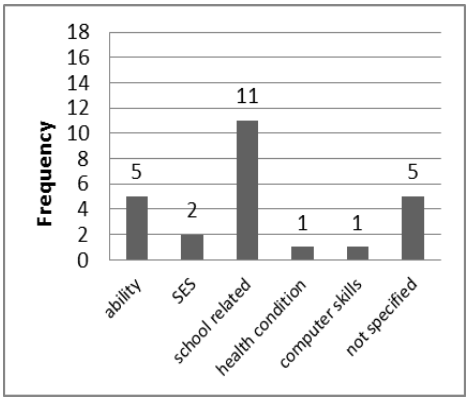

Figure 1. Inclusion criteria (N = 25)

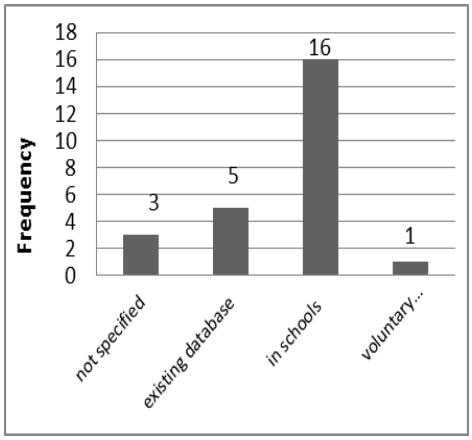

Figure 2. Recruitment of participants (N = 25)

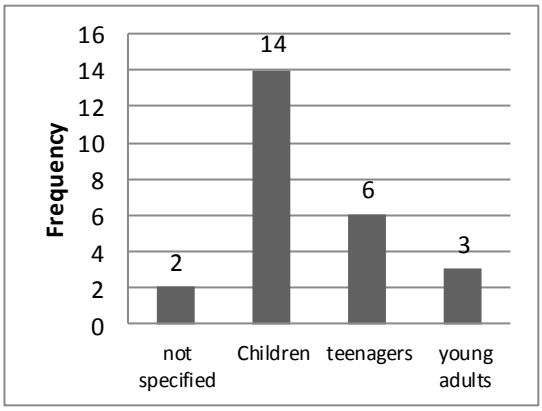

Figure 3. Subjects included in the study (N = 25**)**

Participants were mostly recruited in schools and by using existing databases. One study recruited based on voluntary participation and 3 studies did not specify how participants were recruited (Fig. 2).

The average sample size of participants in studies reviewed was 220 (SD = 284, Mdn = 100), with a minimum of 6 and a maximum of 1274 participants. A minimum of 6 participants spread over several conditions is a remarkably low sample size when carrying out statistical analyses and making certain claims regarding the effectiveness of that particular game, let alone generalizing claims on DGBL effectiveness based on the results of that particular study. 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, Mdn = 46), with a minimum of 2 and a maximum of 758 participants in the experimental and 84 (SD = 92, Mdn = 45) with a minimum of 2 and a maximum of 347 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 (Fig. 3).

5.2. Intervention

In the majority 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 referring to an implementation in a formal institution, such as a school, but where gameplay occurred outside of school hours (Fig. 4). Sixteen per cent did not specify the context of play and 56% did not specify the gameplay composition. 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 for gameplay conditions (Fig. 5). Results of the latter study showed that game play composition influences learning outcomes. More specifically, individual gameplay leads to a significantly better performance. Therefore, 56% studies failing to report on game play composition is problematic.

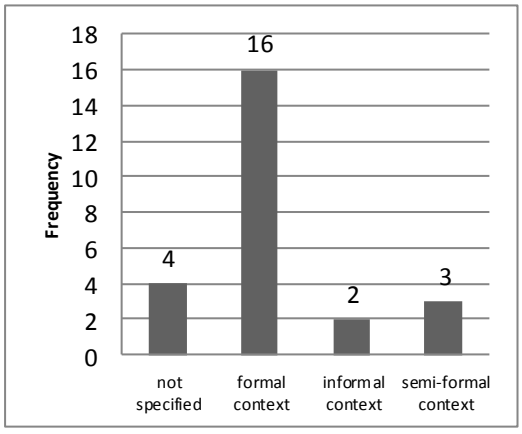

Figure 4. Context of gameplay (N = 25)

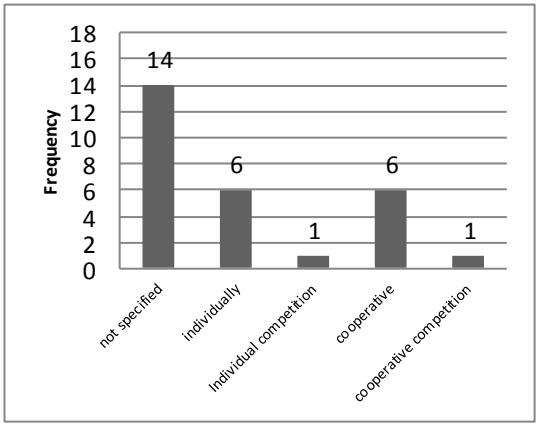

Figure 5. Gameplay composition (N = 25)

Games were either implemented as a stand-alone intervention (28%) or were embedded in a larger program (48%). Forty per cent of the studies did not report on the presence of an instructor, referring to a teacher or researcher present during gameplay. In 56% of the studies an instructor was reported to be present: 5 studies included a teacher as an instructor, 4 studies a researcher, 2 studies university students and 3 studies did not specify the type of instructor present during gameplay. One study did not include any instructor. 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 thus spending additional time with the learning content, leading to confounding effects. Twenty-four per cent did not specify implementation. Table 1 shows an overview of program specifications. While it could be beneficial for DGBL to add elements to the intervention in order to enhance its effectiveness, for the purpose of research aiming at examining whether or not a specific game is effective leads to certain issues. More specifically, these could lead to confounding effects making it impossible for the researcher to know if the positive effects in favor of DGBL were the result of the game as such or the combination of the game with other elements. This is especially problematic when elements containing substantive information regarding the learning content of the game (e.g. extra material, required reading) are added to the DGBL intervention.

Table 1. Specifications of games embedded in a larger program (N = 11)

Program specifications	**N**	**%**	**Description**	**Examples from studies reviewed**
Introduction	5	20	An introduction concerning game content and gameplay was provided by an instructor. This does not refer to an in-game introduction	*…basic instruction in the area of daily economics…Next, the students were shown how to play the game in order to achieve the stated objectives* [24a]
Training of participants before intervention	5	20	A training session before the intervention was provided	*…children were introduced to a 'treasure hunt game' to allow them to develop the skills necessary to navigate in the virtual world of the computer* [3a]
Extra material	8	32	Extra material such as articles, extra exercises, extra reading material, etc. were freely available	*…two classroom instructors, the study guide, their fellow classmates, referenced publications,…* [16a]
Online platform	3	12	The game was part of a larger educational online platform	*Two vocabulary web sites …. Vocabulary games were also available* [25a]
Game task formulation	1	4	Certain tasks were formulated during gameplay	*The students worked together to play the game and synthesize their answers* [24a]

Required reading	2	5	The participants were expected to read next to gameplay	*…required reading for the students were the lab documents…* [1a]
Procedural help by instructor	3	12	The participants received help concerning the actual gameplay. This does not relate to content	*The Computer Science teachers were present in order to provide procedural help to the students, without, however, being actively involved* [15a]
Guidance by instructor	3	12	The participants received guidance during gameplay in order to contextualize the game in the broader learning context	*…instructional discussion between the students and the teacher while the students were playing the game* [5a]
Supplement of course	6	24	Gameplay occurred next to the classes	*After teaching to both groups all required concepts in a regular classroom… a regular set of exercises was given as homework for two weeks to the students of both groups, while students from the test group, apart from the regular exercises interacted with the game during same period of time* [13a]
Debriefing	3	12	A debriefing session was provided	*Once a play event finished, the instructor held a traditional 45-minute 'discussion section' with the students* [17a]

The average implementation period was 9 weeks (SD = 6,7, Mdn = 6), with a minimum of 1 day and a maximum of 23 weeks. Average total interaction time with the game is 12.4 hours (SD = 14.8, M = 9), with a minimum of 30 minutes and a maximum of 64 hours.
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 2 gives an overview of interventions implemented in the control group(s). Thirty-two per cent of the studies reported on how similarity of content in the intervention in the EG and CG was achieved, 24% did not report on this and 12% used dissimilar interventions regarding content. The latter is problematic, considering that in order to make claims on the added value of the DGBL intervention, it should be compared to another educational intervention, covering the same content and preferably instructional techniques with the digital game aspect being the only difference.

Table 2. Interventions in control group (N = 25)

Intervention control group(s)	**N**	**%**	**Description**
Traditional classroom teaching	1 2	48	(One of) the control group(s) got a comparable treatment/intervention by classical classroom teaching

Traditional multimedia classroom teaching	1	4	(One of) the control group(s) got a comparable treatment/intervention by classical classroom teaching with the help of multimedia (video, audio, etc.)
Computer-based learning	4	16	(One of) the control group(s) got a comparable treatment/intervention by a computer-based application, such as an educational website.
Other game	2	8	(One of) the control group(s) got a treatment/intervention by means of another game not related to the subject concerning the game played in the intervention group
Paper and pencil exercises	3	12	One of the control group(s) got a comparable treatment/intervention by means of paper-and-pencil exercises.
No intervention	10	40	(One of) the control group(s) did not get a comparable interventions, bur served as a no-treatment control group.
Not specified	1	4	The study did not report on the type of intervention implemented in the control group(s)

5.3. Method

All studies implemented a quantitative research approach, 32% combined this with qualitative research such as observation, interviews and diaries. However, only 3 studies coded their qualitative data.

All studies reviewed implemented an experimental design. All studies implemented a between-subjects design, with the exception of one study that implemented a within-subjects design, where the game-based group also served as a control group (by implementing traditional classroom teaching before midterm exams and implementing the DGBL intervention before the final exams). Forty-four 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

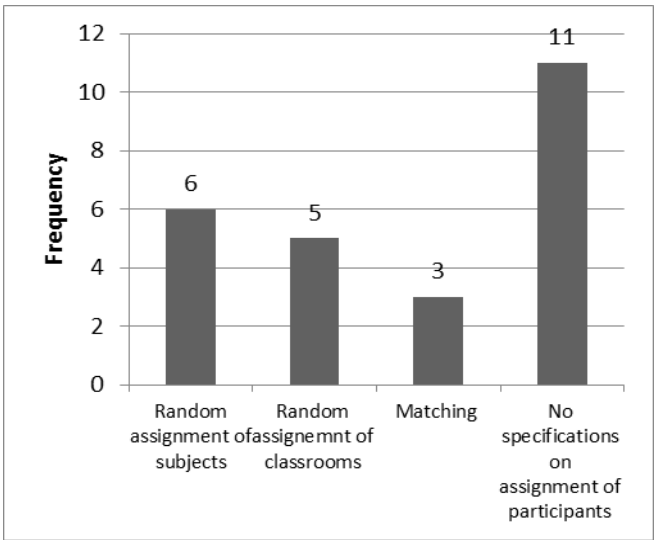

Figure 6. Assignment of participants to conditions

certain characteristics such as previous test scores, and 44% did not specify on group assignment of participants (Fig. 6).

5.4. Measures

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 (8%) used both test scores and student achievement as an accuracy measure. Less than half (44%) implemented standardized tests, six of these (55%) only used standardized tests while 5 studies (45%) combined standardized tests with tests developed by the researchers. 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	%	Description	
Accuracy	19	76		
Test scores	16	64	Absolute test scores of a test developed for the study or a standardized test that has been implemented in the study	The Civics and Society test (CST) was developed using materials provided by the textbook publisher [33]
Student achievement	5	24	Student achievement in the formal context (e.g., exam scores)	…the outcome performance on the midterm examination served as a comparison matched-control, while the outcome performance on the final examination represented the post–digital game-based examination test group. [6a]
Time measurements	2	8		
Time on task	2	8	Time spent on finishing tests	Time taken to complete the challenge was recorded. [34]
Subjective measurements	N	%	Description	
Self-measurements	8	32		
Self-efficacy topic	4	16	Self-efficacy concerning the topic of the game	I'm confident I can understand the basic concepts taught in this course, I believe I will receive an excellent grade in this class [33]

Self-efficacy general	2	8	Self-efficacy on a more general level (e.g., academic achievement)	General academic self [35]
Perceived educational value	2	8	Perceived educational value of the intervention	…questionnaires…in the experimental group in order to evaluate the online resources in terms of their design and effectiveness in helping them learn vocabulary [11]
Motivation	10	40		
Motivation towards educational intervention - Post-only, EG - Post-only, EG and CG - Pre- and post, EG and CG	7 3 2 2	28 12 8 8	Motivation towards learning via a certain intervention	the degree to which they found that the application: (1) was interesting, (2) was enjoyable, (3) was engaging [15a]
Motivation towards learning/educational content - Post-only, EG and CG - Pre-post, EG and CG	3 2 1	12	Motivation towards the actual educational content and not to the way it was delivered	Motivated Strategies for Learning Questionnaire [24a]
Other	2	8		
Attitudes towards school	1	4	Measures for attitudes towards school	…instrument designed to assess children's attitudes toward primary school [12a]
Teacher expectations	1	4	Teachers' expectation of change	In the pretest, teachers must indicate changes expected…In the post-test, teachers must identify positive and negative changes perceived in the dimensions indicated in the pretest… [19a]

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 was 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 for studies a follow-up test. Assessing the lasting effect is, however, important considering that short-term interventions with a new medium can yield a novelty effect, overestimating the instructional value.

Different statistical techniques can be distinguished for quantifying learning outcomes. 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. Table 4 shows how analysis of tests occurred.

Table 4. Data-analysis (N = 25)

Data-analysis	Description	N	%	Examples from studies reviewed
Between groups comparison of difference scores	The difference (e.g., gain scores or percentage of improvement) between pre- and post-test scores are calculated and used as dependent variable in a between groups comparison (e.g., anova, t-test)	9	36	…paired-samples t tests were conducted to compare the treatment and control gain scores from pre-test to post-test…[8a]
Absolute test scores comparison	Differences between experimental and control group are calculated separately for the pre-test (controlling for pre-existing differences) and the post-test scores (e.g., anova, t-test).	5	20	…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 [25a]
Pre-test scores as covariate between subjects	Between groups comparison of absolute post-test scores, controlling for initial levels of ability adding pre-test scores as a covariate	4	16	A 2 x 2 between-groups analysis of covariance (ANCOVA) was conducted to assess the effectiveness of the interventions on students' computer memory knowledge. The independent variables were: (a) the type of intervention, which included two levels (gaming application, non-gaming application), and (b) gender. The dependent variable consisted of scores on the post-test CMKT. Students' scores on the pre-test CMKT served as a covariate in this analysis, to control for eventual pre-existing differences between the groups [15a]
Between groups comparison with repeated measures	Interaction between time (pre-test and post-test) and group (EG and CG) are calculated (e.g., mixed Anova)	4	16	The NTPS scores were analyzed using a two-way mixed design ANOVA, in which instructional treatment was a between-subject factor, while measurement occasion was a within-subject

				factor [24a]
Repeated measures within subjects	A repeated measures for pre-test and post-test score are calculated separately for experimental and control group(s)	1	4	Significant gains were found in the games console group for both accuracy and speed of calculations, while results for the two comparison groups were mixed…The comparison groups showed in significant gains in any area of self-perceptions [11a]
Other	Within subjects design: testing whether or not increased upward shift of scores on pre- and post-tests is statistically significant	1	4	Though the means and the highest scores remained similar, the lowest score shifted from 53.06% on midterm examination to 57.84% on final examination (post–digital game based outcome). This increased positive upward shift was statistically significant at P 5 .04 [6a]
Not specified	Results are discussed without describing the data-analysis methods	1	4	/

5.5 Summary

Table 5 gives an overview of the main differences across studies regarding study design. These elements could serve as a foundation for the development of an overarching methodology for assessing effectiveness of DGBL, examining which elements and which ways of execution lead to more reliable and generalizing results on DGBL effectiveness.

Table 5. Summary of main differences across studies

Aspect of study design	**Main differences across studies (N=25)**
Participants	Large variety in sample size
	Reporting on types of people included
Intervention	Activity implemented in control group(s)
	Stand-alone intervention vs. embedment in a larger program
	Variety of elements present in larger program
	Presence of / role of / type of intermediary
Method	Randomization of subjects/classrooms
	Use of matching in different ways for assigning participants to conditions
	Addition of qualitative data
Measures	Different objective measures of performance

	Different self-report measures
	Similarity pre- and post-tests
	Data-analysis techniques

6. Discussion

The present study indicates that comparison of the results of studies and the making of generalizing claims on DGBL effectiveness is difficult as a result of diversity in study designs, some of which are suboptimal.

Variety in study design is a result of three issues. A first issue is that different activities are implemented in the control group(s). The interpretation of the contribution of the intervention to the EG does, however, depend on the activities performed in the CG [9]. Considering that intervention in the CG can influence results and interventions implemented in CG differed across studies, comparison between study results becomes problematic. A second issue regarding variety in study designs is the different measures that are used for assessing effectiveness. While motivation is considered as an important element in DGBL effectiveness, it is not always assessed. When motivation is assessed, the type of motivation measured and timings of measurement differed across studies. The first type of motivation is motivation toward the educational intervention, gauging for engagement and/or enjoyment during game play and is thus a situational component. This is typically related to measuring concepts as enjoyment, fun and immersion. This is, however, somewhat problematic considering this often implies that the motivation for playing games in the context of DGBL is personal motivation or motivation enabled by the game. As mentioned before, engaging in DGBL is mostly the result of external coercion. To become engaged, a player thus firstly needs to be motivated. In turn, to experience enjoyment and immersion, the player needs to be engaged. [36]. A suggestion made by Schønau-Fog and Bjørner [36] in that respect is assessing the desire to continue playing, investigating the basal level of engagement. The second type of motivation, motivation towards learning or the educational content, however, is seen as an outcome of the intervention. Therefore, it would be interesting to use this measure as a proxy for effectiveness of the educational intervention, considering this could point to a higher interest in the content matter. Consequently, a combination of both types of motivation would be recommended. The development of a validated scale for assessing these types of motivation is therefore an interesting venue for further research.

A third issue is that different statistical techniques are used for quantifying learning outcomes, either comparing gain scores of EG and CG, comparing post-test scores of both groups using pre-test scores as a covariate or using a mixed design, looking at the interaction of time (pre- and post-test) and group (EG, CG). Other studies only compared post-scores, after checking whether the EG differed significantly from the CG on the pre-test. There has been previous discussion in the academic field on how to analyze data of a pre-post control group design [37]. While the use of gain scores has been criticized as being less reliable than using raw scores, it can be used under certain conditions (i.e., pre-test and post-test scores do not have equal variances and equal reliability). These scores cannot, however, be correlated with other variables in the sample. A mixed design would lead to the same results as comparing gain scores [38]. According to several authors, an analysis of covariance (ANCOVA) with pre-test scores as a covariate, is a more preferable method [32, 38]. 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 [38].

Suboptimal study designs are a result of confounding variables influencing the results, leading to insecurity about whether or not the effects found can be attributed to the game-based intervention or other elements added to the intervention during implementation. Confounds should therefore be eliminated as much as possible [39]. There are three types of confounding elements that can be distinguished in the DGBL study design. A first possible confound is the addition of elements to the game used for the intervention. The DGBL intervention is either implemented as a stand-alone intervention or is embedded 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 researcher can therefore not know if positive findings are the result of playing the game or the combination of the game with for instance exercises in a textbook, unless this is

added as an additional condition to the study (e.g., game, game + textbook, control). A second possible confound is the presence of an instructor. If an instructor was present, the type of instructor (i.e., researcher, teacher, student) and the role of the instructor (i.e., supervision, procedural help, guidance) differed across studies as well. Having a teacher as an instructor in a study can, however, result in less control and as a result, confounding variables [7, 40]. Moreover, the presence of an instructor can lead to instructor influences. For instance, a study conducted by Brom et al. [7] has shown that significant findings in one experimental group compared to its matched control group could not be found in another experimental group compared to its matched control group due to teacher influences. Further, offering procedural help or guidance can again lead to an overestimation of the instructional effect of the DGBL intervention [41].

A third possible confound are practice effects when the same test is implemented pre- and post-intervention. when taking an achievement/intelligence test for the second time, participants will automatically do better, even if the intervention would not have taken place. According to Crawford et al. [42] 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 at different points in time (e.g., pre- and post-test), these generally tend to be smaller [43]. The studies in the review that used parallel tests pre- and post-intervention did not specify how this similarity was assessed however. An example on how this could be done, can be found in a study conducted by Nuñez Castellar et al. [10] for instance, where similarity of two parallel versions of a test is assessed by providing one half of the participants with version A and the other half with parallel version B in the pre-test and vice versa. Non-significant differences on the pre-test then refer to comparability of version A and B. Other 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.

Lastly, there are also replication issues with certain studies due to missing information on multiple areas of the study. A detailed description of the procedure is necessary in order to provide other researchers the opportunity to falsify obtained results [44]. Most information is missing on implementation of the intervention, sampling, similarity of the different interventions when other educational interventions are implemented in the control group(s) and information on the tests implemented. The latter two also bring doubt to the validity of certain study results. Information on how similarity between different conditions is attained, is necessary for the reader of an academic publication to know whether different groups were treated in the exact same way with the manipulation (e.g., DGBL intervention) being the only difference considering this is a prerequisite for making conclusions on the effect of the manipulation [39]. Creating comparable conditions is, however, a challenge considering that comparing interactive media content in a game with for instance an oral class given by a teacher is difficult. A suggestion made by Clark [2] in that respect is the implementation of similar instructional techniques (e.g. drill and practice, scaffolding) in the control condition. Consequently, differences in learning outcomes can be attributed to the added value of the medium.

Missing information on the tests that are implemented is also problematic, considering that a general problem in this research area seems to be that test development does not always happen thoroughly enough, again raising questions on their validity [7, 40]. When a test is developed by the researchers, little information is provided on the instruments. For instance, there is often no information on whether or not these tests were piloted. This is important information to provide, however, considering educational research requires rigorous standards of reliability and validity, implying that tests developed by researchers should be piloted and include checks on their internal consistency [25]. Further, objective tests, subjective tests or a combination of both are used for assessing learning outcomes. The mere use of subjective tests such as self-efficacy is considered as problematic, considering student opinion on learning has previously been found to be unreliable and conflicting with direct measures, questioning their validity [2].

7. Limitations and future research

The selection and coding of publications was conducted by one researcher, which can be considered a limitation of this study. This study is also 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 behavioral or attitudinal change.
An interesting area for future research is exploring the possibilities for the development of an overarching methodology to measure effectiveness of DGBL. Further research should therefore firstly focus on the development of an evaluation framework for assessing effectiveness of DGBL in order to develop a common methodology. To be able to develop this evaluation framework, a clear definition of effectiveness in the context of DGBL should be formulated. Considering that there are a lot of stakeholders involved in this field (e.g., game designers, game researchers, adopters and governmental institutions providing funding), this definition should not solely be based on literature reviews, but should also include the conceptualization of effectiveness by these different stakeholders. Moreover, both relevant stakeholders and experts in the methodology field (i.e., educational research and experimental methodology) should be involved in the development of a common methodology in order to find a balance between an ideal research design in terms of validity and what is practically possible.

Lastly, some issues have been raised on confounding elements by implementing the game in a larger program. Empirical evidence on the possible impact of these elements in the context of DGBL research is, to the best of our knowledge, scarce. Therefore, further research on the impact of several factors such as support by intermediaries, program elements and extra material provided, is required.

Acknowledgments

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

References

[1] Sawyer, B. and P. Smith, "Taxonomy for Serious Games". Digitalmil, Inc& Serious Games Initiative/Univ. of Central Florida, RETRO Lab, 2008.

[2] Clark, R., "Learning from serious games? Arguments, evidence, and research suggestions". Educational Technology, 2007. 47(3): p. 56-59.

[3] Mayer, I., et al., "A Brief Methodology for Researching and Evaluating Serious Games and Game-Based Learning". In Psychology, Pedagogy, and Assessment in Serious Games, T.C.T.H.E.B.G.B.P. Moreno-Ger, Editor. 2013, ICI Global.

[4] Mayer, R., "Multimedia Learning and Games", In Computer Games and Instruction, S. Tobias, Editor. 2011, Information Age Publishing: Charlotte, NC. p. 281-306.

[5] O'Neil, H.F.,Wainess, R, Baker E.L., "Classification of learning outcomes: Evidence from the computer games literature". The Curriculum Journal, 2005. 16(4): p. 455-474.

[6] Ke, F., "A qualitative meta-analysis of computer games as learning tools. Handbook of research on effective electronic gaming in education", Hershey: IGI, 2009. 1: p. 1-32.

[7] Brom, C., et al., "Turning high-schools into laboratories? Lessons learnt from studies of instructional effectiveness of digital games in the curricular schooling system", in E-Learning and Games for Training, Education, Health and Sports, Berlin-Heidelberg: Springer, 2012, p. 41-53.

[8] Hays, R.T., "The effectiveness of instructional games: a literature review and discussion". 2005.

[9] Stewart, J., et al., "The Potential of Digital Games for Empowerment and Social Inclusion of Groups at Risk of Social and Economic Exclusion: Evidence and Opportunity for Policy", Institute for Prospective and Technological Studies, Joint Research Centre, 2013.

[10] Nunez Castellar, E., et al., "Improving arithmetic skills through gameplay: assessment of the effectiveness of an educational game in terms of cognitive and affective learning outcomes". Information sciences, 246, 19-31, 2013.

[11] Yip, F.W.M. and. Kwan A.C.M, "Online vocabulary games as a tool for teaching and learning English vocabularyù2. Educational Media International, 43(3): p. 233-249, 2006.

[12] Kretschmann, R., "Digital Sport-Management Games and Their Contribution to Prospective Sport-Managers' Competence Development". Advances in Physical Education,. 2(4): p. 179-186, 2012.

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

[14] Neys, J., et al., "Poverty is not a game: behavioral changes and long term effects after playing PING". In 13th annual conference on the International Speech Communication Association. Portland, 2012.

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

[16] Korteling, J.E., et al., "Transfer of Gaming: transfer of training in serious gaming": TNO innovation for life, 2011.

[17] Gunter, G.A., Kenny, R.F., and Vick, E.H., "A case for a formal design paradigm for serious games." The Journal of the International Digital Media and Arts Association, 3(1): p. 93-105, 2006.

[18] Kozma, R.B., "Will Media Influence Learning? Reframing the Debate". Educational Technology Research and Development, 42(2): p. 7-19, 1994.

[19] Ryan, R.M. and Deci, E.L., "Intrinsic and extrinsic motivations: Classic definitions and new directions.", Contemporary educational psychology,. 25(1): p. 54-67, 2000.

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

[21] Ryan, R.M. and Deci, E.L., "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.", American psychologist, 2000. 55(1): p. 68.

[22] Wilkes, M. and Bligh, J., "Evaluating educational interventions.", BMJ: British Medical Journal,. 318 (7193): p. 1269, 1999.

[23] Calder, J., "Programme evaluation and quality: A comprehensive guide to setting up an evaluation system": London: Routledge, 2013.

[24] Taras, M., "Assessment–summative and formative–some theoretical reflections". British Journal of Educational Studies, 53(4): p. 466-478, 2005.

[25] Hutchinson, L., "Evaluating and researching the effectiveness of educational interventions.", BMJ: British Medical Journal, 318 (7193): p. 1267, 1999.

[26] Hainey, T., "Using Games-Based Learning to Teach Requirements Collection and Analysis at Tertiary Education Level", PhD thesis, University of the West of Scotland, 2010.

[27] Wouters, P., van der Spek, E. and Van Oostendorp, H., "Current practices in serious game research: A review from a learning outcomes perspective.", Games-based learning advancements for multi-sensory human computer interfaces: techniques and effective practices, Hershey: IGI, p. 232-250, 2009.

[28] Shute, V.J., Rieber, L. and Van Eck R., "Games... and... learning. Trends and issues in instructional design and technology", 3, 2011.

[29] Shute, V.J., "Stealth assessment in computer-based games to support learning. Computer games and instruction", 55(2): p. 503-524, 2011.

[30] Higgins, J.P., Green, S. and Collaboration, C., "Cochrane handbook for systematic reviews of interventions". Vol. 5,. Wiley Online Library, 2008.

[31] Kraiger, K., Ford, J.K., and. Salas E, "Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation.", Journal of applied psychology, 78(2): p. 311, 1993.

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

[33] Yang, Y.-T.C., "Building virtual cities, inspiring intelligent citizens: Digital games for developing students' problem solving and learning motivation". Computers & Education, 59(2): p. 365-377, 2012.

[34] Miller, D.J. and Robertson, D.P., "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): p. 242-255, 2010.

[35] Miller, D.J. and Robertson, D.P., "Educational benefits of using game consoles in a primary classroom: A randomised controlled trial.", British Journal of Educational Technology, 42(5): p. 850-864, 2011.

[36] Schønau-Fog, H. and Bjørner, T., "'Sure, I Would Like to Continue" A Method for Mapping the Experience of Engagement in Video Games.", Bulletin of Science, Technology & Society,. 32(5): p. 405-412, 2012.

[37] Singer, J.D. and. Willett, J.B., "Applied longitudinal data analysis: Modeling change and event occurrence.", Oxford university press, 2003.

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

[39] Leary, M.R., "Introduction to behavioral research methods.", Brooks/Cole Pacific Grove, CA, 1995.

[40] Serrano-Laguna, Á., et al., "Learning Analytics and Educational Games: Lessons Learned from Practical Experience.", In Games and Learning Alliance Conference. Paris., 2013.

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

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

[43] Anastasi, A., "Differential psychology: Individual and group differences in behavior.", London: Macmillan, 1961.

[44] Popper, K., "Science: conjectures and refutations. Readings in the Philosophy of Science: From Positivism to Postmodernism", p. 9-13, 2000.

Appendix: Studies included in literature review.

[1a] 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.

[2a] 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.

[3a] 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.

[4a] Din, F. S. and Calao, J., 2001, "The effects of playing educational video games in kindergarten achievement.", Child Study Journal, 31(2), 95-102.

[5a] 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.

[6a] Kanthan, R. and Senger, J.-L., 2011, "The Impact of Specially Designed Digital Games-Based Learning in Undergraduate Pathology and Medical Education", Arch Pathol Lab Med, 135, 135-142.

[7a] 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.

[8a] 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.

[9a] Ketamo, H., 2003, "An Adaptive Geometry Game for Handheld Devices." Educational Technology & Society, 6(1), 83-94.

[10a] 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.

[11a] 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.

[12a] 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.

[13a] Moreno, J., 2012, "Digital Competition Game to Improve Programming Skills." Educational Technology & Society, 15(3), 288-297.

[14a] Moshirnia, A., 2007, "The Educational Potential of Modified Video Games.", Issues in Informing Science and Information Technology, 4, 511-521.

[15a] 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.

[16a] 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.

[17a] 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.

[18a] 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.

[19a] 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.

[20a] St Clair Thompson, H., et al. 2010. "Improving children's working memory and classroom performance", Educational Psychology, 30(2), 203-219.

[21a] 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.

[22a] 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.

[23a] 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.

[24a] 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.

[25a] 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.