Computers & Education 92-93 (2016) 90–103

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Assessing the effectiveness of digital game-based learning: Best practices

Anissa All[*], Elena Patricia Nuñez Castellar, Jan Van Looy

Department of Communication Sciences, iMinds-MICT-Ghent University, Belgium

ARTICLE INFO

Article history:
Received 22 June 2015
Received in revised form 6 October 2015
Accepted 7 October 2015
Available online 22 October 2015

Keywords:
Digital game-based learning
Evaluation methodologies
Evaluation of CAL systems
Interactive learning environments
Media in education

ABSTRACT

In recent years, research into the effectiveness of digital game-based learning (DGBL) has increased. However, a large heterogeneity in methods for assessing the effectiveness of DGBL exist, leading to questions regarding reliability and validity of certain methods. This has resulted in the need for a scientific basis to conduct this type of research, providing procedures, frameworks and methods that can be validated. The present study is part of a larger systematic process towards the development of a standardized procedure for conducting DGBL effectiveness studies. In a first phase, the variety in methods that are used for sampling, implementation of the interventions, measures and data analysis were mapped in a systematic literature review using Cochrane guidelines. The present paper reflects the second stage, where this variety in elements are presented to experts in psychology and pedagogy by means of semi-structured interviews, in order to define preferred methods for conducting DGBL effectiveness studies. The interview was structured according to five dimensions that were used in the literature review: 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 (sampling, assignment of participants to conditions, number of testing moments) 4) outcome measures (e.g., instruments used to measure a certain outcome) and 5) data-analysis. The interviews were transcribed and analyzed using qualitative software package nVivo. Our results show that areas for improvement involve the intervention dimension and the methods dimension. The proposed improvements relate to implementation of the interventions in both the experimental and control group, determining which elements are preferably omitted during the intervention (such as guidance by the instructor, extra elements that consist of substantive information) and which elements would be aloud (e.g., procedural help, training session). Also, variables on which similarity between experimental and control condition should be attained were determined (e.g., time exposed to intervention, instructor, day of the week). With regard to the methods dimension, proposed improvements relate to assignment of participants to conditions (e.g., variables to take into account when using blocked randomized design), general design (e.g. necessity of a pre-test and control group) test development (e.g., develop and pilot parallel tests) and testing moments (e.g., follow up after minimum 2 weeks). In sum, the present paper provides best practices that cover all aspects of the study design and consist of game specific elements. While several suggestions have been previously made regarding research design of DGBL effectiveness studies, these often do not cover all aspects of the research design. Hence, the results of this study can be seen as a base for a more systematic approach, which can be validated in the future in order to develop a standardized

[*] Corresponding author. Korte Meer 7-9-11, 9000 Gent, Belgium.
E-mail address: anissa.all@ugent.be (A. All).

http://dx.doi.org/10.1016/j.compedu.2015.10.007

procedure for assessing the effectiveness of DGBL that can be applied flexibly across different contexts.

1. Introduction

Digital games encompass a variety of types and genres that can be played using a multitude of digital technologies such as computers, (handheld) consoles and mobile devices. Based on a literature review on digital games definitions, Juul (2003) defines a digital game as

> … a rule-based formal system with a variable and quantifiable outcome, where different outcomes are assigned different values, the player exerts effort in order to influence the outcome, the player feels attached to the outcome, and the consequences of the activity are optional and negotiable (p.5).

Digital game-based learning (DGBL) refers to the usage of the entertaining power of digital games to serve an educational purpose (Prensky, 2001). DGBL is the consequence of a balance between learning and gaming elements (Nussbaum & Beserra, 2014). DGBL contains two important elements: fun/entertainment and an educational component (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013). Consequently, in the DGBL literature and published effectiveness studies both learning and player engagement/motivation are considered relevant to assess (Bellotti et al., 2013).

Two types of games can be distinguished in DGBL: special purpose games which have been developed with an educational purpose and Commercial-Off-The-Shelf games that have been developed for entertainment purposes, but that are being deployed in an educational context. Note, however, that this does not mean that special-purpose DGBL games cannot be commercially available (Stewart et al., 2013).

Based on the projected primary learning outcomes, three types of special-purpose games can be distinguished. They aim to achieve knowledge transfer (cognitive learning outcomes), skill acquisition (skill-based learning outcomes), and/or attitudinal/behavioral change (affective learning outcomes) (Stewart et al., 2013). Games that are developed with the primary aim of achieving knowledge transfer are typically implemented in education, in order to teach math (Castellar, All, de Marez, & Van Looy, 2015) or language (Yip & Kwan, 2006), for example. Digital games that primarily aim to support skill acquisition are generally used for training, for example in a corporate or military context. For instance, several studies have examined the impact of playing games to develop managerial skills (Corsi et al., 2006; Kretschmann, 2012). Games that are developed to achieve attitudinal change are sometimes used by governments and NGOs to raise awareness about a certain topic, such as poverty (Neys, Van Looy, De grove, & Jansz, 2012). Games with a behavioral change intention are typically found in the health sector. For example, some games promote healthy food and physical activity to children (Baranowski, Buday, Thompson, & Baranowski, 2008). While DGBL can primarily aim to achieve a certain type of learning outcome, this does not exclude secondary learning outcomes (Kraiger, Ford, & Salas, 1993). For instance, a game that primarily aims to teach children English (cognitive learning outcomes) can also result in a more positive attitude towards learning English or English as a subject (affective learning outcomes).

Although meta-analyses have proven the effectiveness of DGBL (Backlund & Hendrix, 2013; Clark, Tanner-Smith, & Killingsworth, 2015; Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012), certain authors have pointed out elements that jeopardize reliability and validity of some findings. This includes comparisons with control groups that did not receive an educational intervention (Hays, 2005), time-on-task differences between experimental and control groups, and validity of research instruments (Randel, Morris, Wetzel, & Whitehill, 1992). Moreover, some studies do not provide enough information about the implementation of the intervention (Clark et al., 2015; Sitzmann, 2011). This makes it hard for readers to know if the reported results are a consequence of the different methods, and not a cause of circumstantial factors that differed between conditions (Randel et al., 1992). Rigorous assessment is required to improve the quality of DGBL, to support resource allocation, and to gain insight in the most effective way to use games to support learning (De Freitas, 2006; Kirriemur, 2004).

1.1. Studies about DGBL effectiveness

Two types of evaluation of educational interventions can be distinguished. A first type is formative evaluation which aims to determine areas for improvement and is thus an evaluation of the process of the intervention itself. This type of evaluation is conducted by using a naturalistic design with observational data collection, which describes an ongoing process in its natural setting. A second type is summative evaluation, which aims at to determine whether or not an educational intervention succeeds in attaining its goals, thus evaluating the outcomes (Calder, 2013). Summative evaluations are conducted by using an experimental design (Hutchinson, 1999). In the present study, we focus on summative evaluation and will concordantly discuss experimental design.

An earlier content analysis on the effectiveness of DGBL approaches, conducted by the current authors, showed that there is a large diversity in the way that experimental research on DGBL effectiveness assessment is conducted, making comparison of results across studies difficult. This heterogeneity can be found on all four dimensions of the study design, as defined by

Cochrane guidelines, which were used for the content analysis (i.e., a systematic review method which has its origins in health research and aims to assess the effectiveness of interventions for prevention, treatment and rehabilitation (Higgins, Green, & Collaboration, C., 2008). The dimensions are 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 outcomes). Variety is caused by three main issues: the type of activity implemented in the control group (no activity, traditional classroom teaching, computer-based learning, other game, etc.), the outcome measures that are used to assess effectiveness (perceived learning, time on task, test scores, student achievement, etc.), and different statistical techniques that are used to quantify learning outcomes (percentage of improvement, between group comparison with repeated measures, post-test scores comparison, etc.) (All, Castellar, & Van Looy, 2014). Table 1 provides a more detailed overview of the main differences between studies on DGBL effectiveness.

Results of the content analysis also revealed certain suboptimal study designs which are related to confounding elements. Three main issues can be distinguished. Firstly, the addition of elements to the game, such as required reading, extra exercises, or debriefing sessions, makes it impossible to isolate the effect of the game. Secondly, the type of instructor present during the intervention (familiar vs. unfamiliar person) and the role the instructor has during the intervention differs across studies. Instructors are either present to 1) only supervise, 2) offer technology oriented support when respondents encounter issues concerning the technology or actual game play (i.e., procedural help), or 3) offer content-related help, by providing contextualization of game play and in game elements in the broader learning context during actual game play (i.e., guidance) (All et al., 2014). Thirdly, implementation of the same test pre- and post-intervention on the same day, could lead to practice effects and pre-test sensitization. This would, again, result in an overestimation of the instructional effect (Crawford, Stewart, & Moore, 1989; Van Engelenburg, 1999). In 1992, Randel mentioned similar issues with regards to the reliability and validity of certain effectiveness studies on instructional games. Twelve years later, the same issues are still detected in DGBL effectiveness research.

1.2. Towards an overarching methodology

The heterogeneity in study designs, which leads to mixed results and critiques on certain study characteristics, has resulted in a research field which is unable to make generalized claims about the successfulness of DGBL (Giessen, 2015). An underlying reason for this is that DGBL is an emerging field, which combines different disciplines with specific research traditions (Kirriemuir & McFarlane, 2004; Mayer et al., 2014). More specifically, evaluation of DGBL effectiveness is at the crossroads of psychology and pedagogy (Connolly, 2014). Hence, there is a need for an overarching methodology to research and evaluate DGBL, which should provide procedures, frameworks, and methods that can be validated (Mayer et al. 2014). While several suggestions have been made to improve the design of DGBL effectiveness studies (Mayer et al. 2014; Serrano-Laguna et al., 2013), these do not cover all aspects of the experimental research design (e.g., aspects for which similarity between subjects should be attained, instructor role, etc.).

A common methodology would firstly create the opportunity to compare results, and thus the quality, of the different instructional methods across studies. Secondly, claims regarding the effectiveness of DGBL could be made on a more generalized level: per field (e.g., science, math, language learning) or per game genre. Thirdly, a common methodology would set a baseline for quality, which could serve as an evaluation tool for published studies and as a starting point for researchers wanting to conduct a DGBL effectiveness study. Lastly, interest in studying game design features (e.g., competition, narrative, etc.) that influence effectiveness, is growing in order to optimize DGBL game design. In order to make general claims about game design features that influence effectiveness, a standardized approach for studying effectiveness is required (Cagiltay, Ozcelik, & Ozcelik, 2015; Kirriemuir & McFarlane, 2004).

Table 1
Main differences across DGBL effectiveness studies regarding methodology.

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	• Assignment of subjects to conditions • Use of matching/blocked randomized design in different ways • Addition of qualitative data
Measures	• Different objective measures of performance • Different self-report measures • Similarity pre- and post-test(s) • Data-analysis techniques

(Adapted from All et al., 2014, p. 14).

As the field of DGBL effectiveness research is relatively new and no guidelines currently exist for conducting these types of studies, expert interviews are considered appropriate for defining best practices. Expert interviews are typically used for exploration in an emerging field and allow the accumulation of both process and context knowledge (Flick, 2009). Experts possess both theoretical knowledge and experience with actually executing experimental research. Thus, they know how to tackle some issues related to experimental research in a DGBL context. The present paper aims to formulate best practices based on expertise coming from both experimental research in both psychology and pedagogy in order to create a more standardized evaluation approach.

In the present study, as part of a larger process towards a more standardized approach for conducting DGBL effectiveness studies, we focus specifically on special purpose DGBL. Considering the different types of learning outcomes require different types of assessment (Kraiger et al., 1993) and thus resist categorization in one research taxonomy, we solely focus on cognitive learning outcomes.

2. Method

Experts have been defined as *'staff members of an organization with a specific professional function and a specific experience and knowledge for this purpose'* (Flick, 2009 p.166). Based on this description we selected professionals with at least a Ph.D. degree in either educational sciences or psychology who have conducted, or are still conducting, relevant research which evaluates educational interventions. We used a combination of purposive sampling, based on the criteria stated above, and snowballing (Flick, 2011). Interviewed experts were requested to provide other experts who could be relevant for this study. Seven experts in psychology (five national and two international experts) and six experts in pedagogy (two national and four international experts) were interviewed. Ten interviews were conducted using videoconferencing software and three were conducted face-to-face.

Interview questions were derived from a review study conducted by the authors (see paragraph 1.1.). The interview was structured according to four dimensions of Cochrane guidelines that were used in the literature review: 1) participants 2) interventions 3) methods and 4) outcome measures. We have added a fifth dimension, data-analysis, as the content analysis of the authors indicated that different statistical techniques are used to quantify learning outcomes in DGBL effectiveness studies (All et al., 2014). Hence, the interview was conducted according to five instead of four dimensions. The interview guide can be found in Appendix A.

The interviews were conducted over a period of two months. During this period, interviews were transcribed and analyzed using the qualitative analysis software package nVivo. The transcribing and coding did not occur at the end of the process. Instead, the 'constant comparison' principle was applied: this refers to simultaneous relationship between collection and analysis of data (i.e., not a sequential relationship where all interviews are conducted first followed by analysis of all these interviews) (Glaser & Strauss, 2009; Suddaby, 2006). This allowed us to conduct interviews until 'category saturation' was achieved (Strauss & Corbin, 1990). Signals of saturation are the '*repetition of information and confirmation of existing categories*' (Suddaby, 2006, p. 639). Hence, data collection stopped when no new codes were developed during the analysis. Thirteen semi-structured interviews were conducted in total, which is considered an acceptable sample size for expert interviews (Baker & Edwards, 2012).

The interviews were analyzed inductively and the experts' answers and comments were coded in three phases (Thomas, 2006). In a first phase, the transcriptions were coded at the lowest level. This means segments of texts were labeled using in vivo coding (i.e., defining the text by a concept/phrase used by the interviewee). In a second coding phase, labels referring to similar content were grouped and conceptualized, creating categories (e.g., similarity between conditions). In the last phase, these categories were attributed to dimensions of the study design (e.g., research design, participants, intervention, outcome measures and data analysis). An example of how coding occurred can be found in Table 2.

3. Results

3.1. Research design

3.1.1. Control group

Experts were asked what the most ideal design would be, when assessing the effectiveness of a DGBL intervention aimed to achieve cognitive learning outcomes. The majority (11/13) of the experts expressed the need for a control group. This would allow evaluation of two aspects: (1) if positive outcomes are related to the mere lapse of time and (2) the comparison of motivational outcomes.

Two experts, however, stated that there is no need for a control group. According to them, determining whether or not predefined goals are attained would be sufficient. The reason for this is that firstly, a control group is only required when the research question involves a comparison with another method. Secondly, many differences can exist between the educational intervention the game is compared to and the game itself, resulting in a flawed comparison.

It was, however, generally agreed that when a control group is included in the research design, another educational activity should be implemented in the control group. Comparing with a control group that does not receive instruction only results in knowing whether being exposed to certain content through DGBL games is better than not being exposed to any learning content at all, which does not add value to either the research field or society. It would be most preferable to compare

Table 2
Example of how transcriptions were coded.

Phase 3 (dimensions of the study design)	Phase 2 (creating categories)	Phase 1 (in vivo coding)	Example of quote
Intervention	Similarity between experimental and control group	Equal time exposed to intervention	All these interventions should however be matched on different criteria, for instance: content, time of exposure
		Same content in conditions	If you have more difficult questions on the paper and pen exercises; then yeah … the content should in principle be the same. The subject of study is a medium, so the content in both should be identical.
		Same interaction with other people across conditions	The amount of interaction one has with other people, I do not know if this is always matched in both conditions. That seems relevant to me.
		Support received	If you think as an instructor that this intervention requires explanation, above what you would usually give if you use the conventional method, this might trigger better learning. But had you given the same explanation to the conventional method, the effect would have been the same, right? So, there have to be rules on how much you reveal or how much you help people because it's a new environment and this level of help and support and assistance has to be comparable across this two treatments. And I agree that the game methods, the educational game treatment will require maybe some more help. But it has to be specified very clearly why you use help, why you intervene, and if you do, it would be important to do it for everybody. Right?
		Same instructor across conditions	… whether it was the same instructor who instructed the people in the game group and in the control group because it might be an instructor effect if these are different people, right?

with ‘business as usual’. In the context of educational research this could, for instance, be traditional classroom teaching, doing exercises, or the use of another electronic platform (if this is how pupils are currently being instructed, for example in long-distance education). The implementation of another digital game in the control group, not related to the learning content, would only be interesting when examining the motivational rather than the cognitive aspects of DGBL. One expert stated that another game could also be implemented as a third condition.

> That way, you can examine whether or not the combination game and math is of added value. If you for example find higher results for condition 1 (combined condition game + learning content), compared to condition 2 (learning content in a traditional way) and condition 3 (game without learning content), you know the combination of the two elements is superior. If you only find a difference between condition 1 and condition 2, it can be attributed to the fact that it was a game, and therefore participants were more motivated. And if you only look at the difference between condition 1 and condition 3, you could say that it is due to the integrated learning content in the game. While if you find this result, you know the combination has an added value. This is a conclusion you cannot reach while only working with condition 2, or condition 3 compared to condition 1.
>
> (Experimental psychologist, Professor in experimental methods, Belgium).

One expert stated that the comparison with another, non-digital game would be of interest, considering these entail similar processes.

> With hindsight, I now favor the idea of having - and one of my doctoral students has just done a study using this approach - the controls having a parallel, if you like, or a similar learning experience which addresses the same learning outcomes and the same processes, but not using technology. In fact, she [doctoral student] had the children using card games. It is just another interesting way of looking at it: any differences can then be attributed not to extra learning or the introduction of new concepts, but to the actual technology.
>
> (Educational scientist, Professor in educational studies and research methods, Scotland).

3.1.2. Pre-test

A pre-test was generally considered indispensable in this type of research for three reasons. Firstly, when no pre-test has been implemented, differences between the experimental and control group at the beginning of the intervention are not controlled. This means that the experimental group could, by chance, have had higher levels of knowledge before the intervention, resulting in an overestimation of the effect of the game-based learning intervention. Secondly, a pre-test is necessary in order to determine the relative learning gains of the participants as a result of the intervention. Thirdly, when no pre-test is implemented, there is no control for characteristics of drop-outs, which is especially relevant in the context of educational research.

> What you often see in this type of studies, is that drop-out is not random; especially participants who performed poorly, don't feel like participating anymore. They might find it confrontational to come back, because they feel ashamed.
>
> (Experimental Psychologist, Professor in data-analysis and statistics, Belgium).

Two experts also stated that the ideal design would be a Solomon 4-group design, creating four conditions, of which two conditions -one in the experimental and one in the control group-do not receive a pre-test, in order to control for practice effects (i.e., when taking the same test twice, participants generally do better the second time). One expert did, however, state that a pre-test is not always necessary, if the learning goals of the intervention are clearly defined. When no pre-test is administered, randomization of subjects should be applied.

3.1.3. *Follow-up study*

Furthermore, the vast majority (11/13) of the experts considered the integration of a follow-up study as good practice. This is especially relevant with short interventions, in order to examine whether or not the effect was a result of intensive training. Several experts indicated that in educational research, effects that disappear after a few days have little use. A minimum 'longer term' assessment required would be two weeks. Ideally, the period would be three to six months and up to one year for longer interventions. Moreover, instead of announcing the follow-up study, a surprise recall would increase ecological validity. The experts, however, are aware of the fact that organizing a follow-up study after, for instance, one year is rather difficult in practice due to attrition (i.e., loss of cases over time).

3.1.4. *Assignment of participants to conditions*

Randomization has been accepted by all experts as the preferred method in order to keep groups as similar as possible in terms of gender, age, motivation, etc. Two types of randomization were discussed: randomization of subjects and randomization of classrooms. While experts mention randomization of subjects as the most preferable method, they acknowledge that this is not always possible in real life for two reasons. Firstly, randomization of subjects entails the need for a larger sample size, which is often an issue in this type of research. Secondly, randomization of subjects is not always possible due to the context of the study (e.g., implementation in a natural collective such as a class group). Due to the practical limitations mentioned above, randomization at the classroom level would be another option. A limitation of this type of randomization is that school influences could result in a biased sample.

> The main problem with this [randomization on the classroom level] is that you might bias your sample; if for instance you have a school where only pupils of low socio-economic status go, than if all the classrooms come from this school, results might not be biased. You just have to make sure that you mention that these results are only valid for children with low SES. If you have mixed classrooms, which mix children with a high SES and a low SES, and by chance you only give the manipulation to children with high SES … If you conclude that the manipulation worked, you might have biased results.
>
> (Experimental psychologist, Researcher on digital-game based learning, Belgium).

Furthermore, classroom influences related to teacher characteristics (e.g., teacher style, experience with and attitude towards games) could lead to biased results. Matching has also been suggested by the majority of experts (12/13) as a way of guaranteeing similarity between conditions, controlling for certain variables.

> I think matching on characteristics is better than doing no matching, but it's not as good as randomization. There could be unobserved characteristics that you are unable to match on. So, it could be a matter of student motivation. It could be that only the most motivated and hardworking students are going to sign up for an online course versus a traditional course because it requires more self-motivation on the part of the student to get it done without a teacher there watching them every minute. So you do have to worry about differences in unobserved characteristics if you match on characteristics like test scores. As I mentioned earlier, I have done some other studies that have used matching and I think it's better than not doing any type of control for comparison conditions, but it's not as good as a randomized controlled trial.
>
> (Educational Scientist, Senior research scientist, assessment of educational interventions, US).

Table 3 provides an overview of variables to match participants in different conditions as suggested by the experts.

Table 3
Variables suggested to match on.

Variable	Description
Previous knowledge	Matching on prior academic achievement or pre-test scores
Ability	Matching on different ability levels (e.g. low, medium and high achievers)
Motivation	Matching on motivation towards the learning content
Game experience	Matching on previous experience with games
Gender	Matching on gender (male/female)
Age	Matching on age/age categories
SES	Matching on socio-economic status

3.2. Participants

3.2.1. Sample size

An absolute minimum suggested by the experts is 20 participants per condition. For more sophisticated statistical analyses, a required minimum would be 30 participants per condition. Several experts suggested conducting a power calculation beforehand. This would serve as a basis for determining how large the sample size should be in order to detect a real difference (and not miss them). Assumptions would have to be made about the magnitude of the effect (e.g., effect size) to determine power. The calculation itself could be added in an appendix.

3.3. Intervention

3.3.1. Context of the intervention

With regards to context of play, expert opinion is divided. Four experts have a strong preference for implementing DGBL in a formal context. This creates opportunities for more control and should result in a higher internal validity. The other experts have a preference for a context that is representative for the game implemented, in order to increase ecological validity.

> ‘I think that you should describe, as much as possible, how you implemented the game. At least there should be some reference to a website, for instance, where the game and intended gameplay is described, in case there is not enough room in your article.’
>
> (Educational scientist, Professor in research methods and statistics, The Netherlands).

With regards to an informal gameplay context (e.g. at home), certain issues about control are raised. Nevertheless, control should be possible in this context according to one expert.

> ‘I find it no problem that children play a game at home, but you should be sure that they do it under certain conditions, such as under parental supervision, and that you can also define that they play during specific hours or a maximum during a time period for instance.’
>
> (Cognitive psychologist, researcher on digital-game based learning, Belgium).

Logging could also be an opportunity, but by itself it does not ensure control.

> ‘You can log what they do and can keep track of how long they are playing … but still, then you don't really know if they just leave the game ‘open’ and go for a drink in the meantime for instance. So you still have some things you can't control.’
>
> (Educational scientist, Professor in educational technologies, Belgium).

There is a general agreement however, on conducting these types of studies in a lab environment. The lab is a very controlled way of doing research, which will not be representative for the real world. A lab study could, however, serve as a first phase in research.

3.3.2. Similarity intervention and control group(s)

One of the major limitations observed by the experts is a lack of reporting on the similarities between conditions. Ideally, these are the same except for one aspect: the digital game component.

> Apart from the activities implemented, all conditions should be as similar as possible. If you want to examine whether or not learning through a videogame is better than another method, only the game element can be different between the conditions. I definitely think that authors do not report enough on this subject: they do not provide enough information about the conditions, the implementation of the interventions, and how similarity between conditions is attained. In general, I think that the interventions are poorly described. I mean … what do they mean by traditional classroom teaching? I don't know what that is; is it the same as classical classroom teaching? I don't know. For me, it's not the same thing. And isn't this something that is culture dependent? The meaning of traditional classroom teaching will be different in America, Russia, China, etc. So that doesn't say anything.
>
> (Educational scientist, Professor in instructional psychology and technology, Belgium).

Table 4 gives an overview of aspects on which researchers should try to attain similarity between conditions according to the experts.

3.3.3. Instructor

Whilst no preferences regarding the type of instructors present during the intervention were expressed, advantages and disadvantages of the different types were discussed. One disadvantage of adding a non-researcher, such as a teacher, as an instructor is that there is teacher influence: this might impact the results. An advantage of using a researcher as an instructor is that they are trained to give the same instructions to the subjects in each condition in a controlled manner. The presence of a researcher as an instructor can, on the other hand, also lead to bias: people react to changes in environments, for example by aiming to make a good impression. In order to strike a balance between the internal and external validity of a study, experts

Table 4
Aspects of intervention where similarity should be attained in different conditions.

Aspect of intervention	Description
Time exposed	Time exposed to intervention should be exactly the same in both conditions
Content	The exact same learning content should be present in both conditions.
Instructor	The instructor in the different conditions, should be the same person across all conditions.
Support received	Technical support or guidance received by the instructor/intermediaries.
Difficulty level	Content treated in all conditions should be of the same difficulty level
Interaction with other people	Amount if interaction other participants or instructor during the intervention.
Day of the week	Day on which the intervention took place, should be the same in each condition.
Environment	The intervention for all conditions should be conducted in the same environment (e.g. the same classroom)
Types of exercises	Types of exercises used in the game condition should be the same in the other conditions.
Awareness of testing moment	If the game group is briefed about testing moment after the intervention, the other conditions should also be briefed on this.
Reward for participation	If a reward for participation is provided to the game group, this should also be provided to the other conditions.

suggest developing procedures for teacher instruction. In order to control the correct implementation of the procedure, observation by the researcher would be ideal.

In relation to the actual support provided by the instructor, some best practices were mentioned. Whilst some of the experts stated that the most ideal situation would be the absence of an instructor, in order to isolate the game component, even these experts were aware that this is not always possible in this type of research.

> With a good game, the teacher isn't involved at all. In most traditional classroom teaching situations, the teachers are there, at the very least monitoring; offering the odd bit of advice and support. So, it has to do with letting the reader know about the context as a whole.
>
> (Educational scientist, Professor in educational studies and research methods, Scotland).

It is of great importance that the same support is provided in each condition. In some cases, experts prefer that no support is provided at all in order to avoid a confounding effect. Providing procedural help (e.g., support when respondents bump into issues concerning the technology or actual game play or technology oriented support (All et al., 2014) is generally not perceived as problematic: this support would occur even in a formal and even in an informal natural setting (e.g., at home, parents providing this type of help to their children). Consequently, by not providing any support, ecological validity might also be jeopardized. An expert, however, noted that the amount of support that will be provided will be influenced by certain context variables, such as the tech savviness of the parents or teacher or attitude towards games. Hence, it is important to take this into account. Therefore, when it is provided, a clear description is required of what this procedural help actually consisted of.

Guidance (e.g., teacher/supervisor helps to contextualize game play and in game elements in the broader learning context; help related to learning content (All et al., 2014)), however, could lead to problems with internal validity. This raises the question of whether or not an effect would have been found if no guidance was provided. Providing guidance also leads to problems with comparability, considering the game condition might need extra help. The absence of guidance, however, could also lead to problems with ecological validity in certain contexts, such as a classroom environment, since asking questions and offering guidance is generally present in this type of environment.

> There should be some rules about that [role of the instructor during the intervention]. Because, if you, as an instructor, think that this intervention requires explanation beyond what you would usually do with a conventional method, this might trigger better learning. Had you given the same explanation to the conventional method, the effect would have been the same, right? So, there have to be rules on how much you reveal and how much you help people because it's a new environment and this level of help and support and assistance has to be comparable across the two treatment conditions. I agree that the educational game treatment might require some more help. But it has to be specified very clearly why and how you assist, why you intervene, and if you do, it would be important to do it for everybody.
>
> (Experimental psychologist, Researcher on quantitative methodology and experimental methods, UK).

3.3.4. Implementation

One expert believed that DGBL should be implemented as stand-alone intervention. Four experts stated that extra elements could be added to the intervention, as long as these elements are the same in the different conditions. Six experts agreed that the game component should be kept as isolated as possible, considering that other elements added to the intervention could influence learning and thus confound results. Nevertheless, some elements that might be indispensable for practical reasons could be allowed. These elements should, however, be offered in all conditions: for example, even the non-game group should get a training session with the game if this is provided in the game group. The introduction and/or training sessions should not contain substantive information on the learning content covered in the game and should only cover such elements as getting acquainted with the storyline and controls, for example. Otherwise: ‘participants with lower computer skills or less game experience will use up more cognitive capacity in trying to understand the environment, instead of the game or the idea behind the game' (Experimental psychologist, researcher on quantitative methodology and experimental methods, UK).

Several experts also stated that the necessity to provide procedural help, as discussed in the previous section, might be avoided by providing a training session before the intervention, similar to a trial exercise in psychological experiments. According to more than one third of the experts a debriefing session would not be problematic, if implemented after the post-test, considering a debriefing can entail a learning effect.

If one purely wants to assess the effectiveness of the game itself, providing substantive information regarding the content of the game (e.g., required reading, extra material freely available, game task formulation, etc.) might confound the effect. Therefore, such elements are best excluded from the study.

> Elements such as Supplementary material, integration in online platforms, and training sessions result in a confounding effect, caused by extra materials, extra attention, or extra help. This is not the way effectiveness research should be conducted. You should not add extra material if you purely want to assess the effect of the intervention. That is something you should not do. If you want to look at an educational program, of which the game is a part, you can add extra material. But then you should report on the effectiveness of the program and not of the game.
>
> (Educational Scientist, Professor in research methods and statistics, The Netherlands).

Two experts also suggested that, before considering the addition of elements to the digital game, a pilot study on the impact of these different elements of implementation on the learning outcomes could be conducted.

> This has to be pilot tested. If any doubt exists about a training session and the impact it might have on the intervention, that is something you would want to know at the beginning of the study … You can pilot test this, and with this I mean with small groups of 10–15 people. Then you execute the intervention without, for example, an introduction and ask the people afterwards whether or not they thought it helped, if they find it necessary, if they had not received the introduction. This way, you have at least some input on that. Whether you eventually use it in your intervention, is up to the researcher, but it is the total package that counts. The effect will be of the intervention as a whole: afterwards you can only speculate about what it would have been without the introduction; but that is another study.
>
> (Experimental Psychologist, Professor in data-analysis and statistics, Belgium).

3.4. Measures

3.4.1. Instrument validity

It was generally accepted that when a standardized test on the subject covered in the intervention is available, its use is preferred over one developed for the occasion. The self-developed content could be too closely aligned with the intervention and thus bias the outcomes of the study. It was also recognized that finding a suitable standardized test is not always possible.

When developing a new test it is considered important that the process of development is clearly described: who was involved in the development? Was an expert on the content domain involved? Which factors were taken into account? Et cetera. Moreover, the following elements for reporting on tests used were brought forward by the experts: type of knowledge measured (e.g., knowledge, insight, problem solving, creative ability, etc.), type of questions (e.g., open or closed), difficulty level of the questions, and psychometric properties of the scales used.

Moreover, according to one expert, results should contain average scores without any intervention, scores within certain age categories and a normal distribution of the scores in the target population. Furthermore, several experts favor a 'full disclosure' mentality, adding instruments used if possible. If space is limited, authors should indicate where the test can be found. Finally, several experts suggested that a pilot test would increase validity. This could be done by, for instance, implementing the tests in a similar group of participants, conducting cognition interviews while filling out the tests, examining whether the questions are clear and are interpreted the way they are intended.

With regard to the usage of student achievement (e.g., exam scores) as an indicator of learning outcomes, a majority of the experts (9/13) found it a relevant measure, in terms of ecological validity. However, as student achievement scores can be influenced by other factors, it should be combined with more specific content tests.

> The ultimate goal of serious gaming in a school context would be finding an effect on the school scores (i.e., grades on a test or exam) right? So if you find an effect on your post-measures, but cannot find an effect on school scores, you have a problem. So, for a complete assessment, I think you need both.
>
> (Experimental psychologist, Professor in research methods and statistics, Belgium).

One expert also indicated that a prerequisite for using student achievement as a measure would be that participants come from the same school. When several schools and educational levels are included in the study, student achievement measures will become difficult to compare. Finally, one expert stated that student achievement only seems relevant for longer-term interventions that have been introduced over a whole semester, for instance.

3.4.2. Similarity pre- and post-test

When using the same test pre- and post-intervention researchers should be cautious about a practice effect. According to three experts, this is less of an issue if the same test is implemented in the other condition(s) as well, considering the primary interest would be the difference between the conditions. Practice effects also depend on the time between the pre- and post-test. A standardized test would be advantageous here, considering these tests typically have guidelines on a minimum amount of time needed between the pre- and post-test.

Using a similar test (e.g., same type and difficulty level of questions) pre- and post-intervention is considered to be a better option by the majority of the experts (8/13), but similarity of both versions needs to be established. Content experts should be involved in the development, but could also serve as ‘external assessors’. A better option would be to conduct a pilot study with a group of participants who do not receive the intervention, testing every question separately to see if questions of both tests are matched.

> If you run parallel tests, you should test these beforehand. You would have to match the questions that have the same difficulty level; so that the scores are the same when you implement these tests in two groups. You should also examine the average scores and match questions that in the first version yielded, for instance, 50% correct answers to questions that resulted in 50% correct answers in the second version. And then, you have to make sure that all themes are present in the same proportions in each version. This is not always possible because it implies you would have to conduct a large pilot with the tests you know you will use in your actual experiment later on. Tests should already be created long before your actual study, which is sometimes impossible due to timings and so on. So, if you do not have the luxury of time and you have to create tests that are similar, use some of the same questions in both versions and use some parallel questions that have been evaluated by external assessors on similarity, regarding the type of questions and difficulty level.
>
> (Educational Scientist, Researcher on teaching methods, instruction and assessment, Belgium).

3.5. Data analysis

The majority of the experts (10/13) would opt for a standard repeated measures design. This would control for pre-existing differences and take pre-test scores, post-test scores, and comparison of progress across groups into account.

Two experts stated that a mixed effects model should be used, taking both fixed and random effects into account. Fixed effects refer to the effects that are the object of study, which is the instructional condition in this case. Random effects refer to any elements that are observed, and which might lead to extra variance on top of the experimental variance (i.e., the variance between conditions as a result of the difference in treatments, such as different instructors). These are thus potential confounding variables and they should be added as random effects in the model.

> When you conduct analyses, statistics, you have your fixed effects - which you have manipulated – and random effects. In a normal analysis of variance, those random effects could be … you could name thousands, such as sequence, etc. In mixed effect models, those random effects are included in the model.
>
> (Experimental Psychologist, Professor in psychology, Belgium).

Further, experts emphasized the importance of individual differences in educational research. When aiming to control for certain characteristics, an analysis of covariance is preferably used. Characteristics that are to be examined are added as covariates. Interesting control characteristics suggested by the experts are ability (low, medium and high achievers), computer skills, and previous game experience.

> If there is one area where individual differences are very important, it is the learning context. Therefore, the question is not if learning with a game is better than learning without a game. I would preferably put the emphasis on who actually benefits from this game-based approach and who does not. These kinds of individual differences should be taken into account and these variables can be added as a covariate in your analysis. You will notice that there are a lot of differences. This is something you can even do afterwards, without the control group. If you only have the game with its learning effect, you will have people who improve, who worsen, and those where nothing much happens and this could be due to a variety of elements.
>
> (Experimental Psychologist, Professor in psychology, Belgium).

When differences in pre-test scores are found between conditions, pre-test scores could also be introduced as covariates in order to take these pre-existing differences into account in the analysis.

With regard to reporting, several experts stated that description should not only include significance levels: effect sizes should be reported as well for several reasons. Firstly, reporting on effect size is important to make meta-analyses possible. Secondly, effect size provides more information than significance levels; it also shows how large the effect is. Lastly, by using effect sizes a researcher can estimate how much variance is actually explained by the condition and how much variance is explained by other variables, such as attitude or student achievement.

3.6. Overview

Below, on overview of best practices is proposed, based on the expert interviews, which can serve as a guide for the design of future studies (see Table 5).

Table 5
Summary of best practices for effectiveness assessment of DGBL.

Best practices	Advantages	Disadvantages
Research design: general		
1 Control group		
1.1 Control group 1: 'Business as usual'	Control if positive results are not result of mere lapse of time Compare motivational aspects	Larger sample required
1.2 Control group 2: game without educational content (optional)	Compare motivational aspects	Larger sample required
2 Pre-test	• Control for pre-existing differences between experimental and control group(s) • Determine progress/learning gains as a result of the intervention • Make it possible to control for characteristics of drop-outs (i.e., random or non-random attrition)	Practice effects
3. Similarity between experimental and control group should		
a) Randomization of subjects	Balanced groups in terms of observed and unobserved variables	Larger sample required
or		
b) Randomization of classrooms/schools	Often more practical in educational research	Classroom/teacher/school influences
or		
c) Matching (blocked randomized design)	Control for similarity between conditions	Unmatched latent variables can influence results
4 Follow-up study (Min. 2 weeks after intervention has finished)	• Control for novelty effect • Control for positive results as a result of intensive training • Control for longer term effects	Attrition
Intervention		
5 Training session	• Might reduce the need for procedural help during the intervention • Less cognitive load is used up for learning to work with the game-environment	Might bias result
6. DGBL as stand-alone intervention	Potential confounds are reduced	Ecological validity might be reduced
6.1. No adding of elements that contain substantive information		
6.2. Instructor role reduced to procedural help		
7 Instructor type		
a) Researcher	More experimental control	Ecological validity is jeopardized
or		
b) Familiar person (current teacher)	Increases ecological validity	• Less experimental control • Teacher influences
8. Similarity between interventions should be assured by:	Potential confounds are reduced	Ecological validity might be reduced
8.1. Time of exposure		
8.2. Content		
8.3. Support received		
8.4. Environment		
8.5. Awareness of testing moment		
8.6. Reward for participation		
Participants		
9 . Min. 20 participants per condition	Determine differences in dependent variables between groups	More sophisticated analyses are not possible (e.g., covariance adjustment)
Measures		
10 Instrument validity	• Have been validated • Provide suggestions with regard to timespan required between pre- and post-test for administering the same test	• Might not exactly cover what has been discussed in the game/traditional class
10.1 Standardized tests		
or		
10.2 Ad hoc test developed by researcher	• More closely aligned to content treated in game/traditional class	• Pilot study required • Content might be too closely aligned to content treated in game/class
11 Student achievement (e.g., exam scores)	• Ecological validity measure	• Can be influenced by other factors than the game/control intervention • Pupils should come from the same school (or even class) • Only relevant for longer term interventions (e.g., a whole semester)
Data analysis		
12 . Repeated measures	• Analyze interaction between condition and time	• Differences with regard to the dependent variable(s) can exist between groups before the intervention
13 . Add random effects if observed	More precise estimate of the treatment effect	Larger sample size required

Table 5 (*continued*)

Best practices	Advantages	Disadvantages
14 . Add participant characteristics as covariates (e.g., game experience, computer skills, ability)	Take into account individual differences in order to determine for whom DGBL interventions are more beneficial	Larger sample size required

4. Discussion

The present study aimed to define best practices for assessing the effectiveness of DGBL, based on interviews with experts in pedagogy and psychology.

When we compare our results to current practices in DGBL effectiveness assessment, some areas can definitely be improved. Firstly, follow-up studies are rarely implemented (Backlund & Hendrix, 2013). They are considered important in the case of DGBL due to the 'novelty' of the gaming medium (Clark, 2007). Follow up studies help to establish if short-term beneficial effects might be an overestimation of the instructional effect. Secondly, the role of the instructor should be reduced to procedural help in order to define the effectiveness of the game as such. This conflicts with current DGBL literature, where the instructor and the creation of a meaningful learning context are considered key elements in achieving DGBL effectiveness (Cruz, Cruz, Ruiz, David, & Hernández, 2015; De Freitas, 2006; Giessen, 2015). Thirdly, a large part of the present studies have (unintentionally) added confounding elements to the intervention (All et al., 2014), making it impossible to know whether or not an effect would have taken place without these elements. Future studies should aim to isolate the game as much as possible and provide more information on characteristics of the implementation of interventions. Lastly, more information should be provided on efforts to achieve similarity between experimental and control conditions, in order for readers to judge the quality of the papers.

The present manuscript confirms several general best practices in experimental research on educational interventions that should also be implemented in DGBL effectiveness research. However, it also brings forward several best practices specific for experimental research in a DGBL context. For example, individual characteristics/differences are more important compared to 'general effectiveness evaluations' when conducting DGBL research. Previous game-experience and computer skills might also influence the effectiveness of the DGBL intervention. Hence, these variables are important to take into account and control for when assigning participants to conditions.

Beyond the characteristics of the learners themselves, the characteristics of the environment in which DGBL is implemented are relevant. For instance, if implemented among children at home, tech savviness or parental gaming experience can play an important role, as it might lead to increased procedural help (help related to issues regarding technology or game play). If implemented in a school context, tech savviness of the teacher and his/her attitude towards digital games as an instructional tool could also influence results, and this should be taken into account. Moreover, DGBL effectiveness studies in a lab context should be avoided, due to the important role of intrinsic motivation/enjoyment plays in the learning process. This is why it is important – besides the general good practice reasons – that a control group receiving another educational activity is added to the study design (i.e., in order to compare motivational outcomes). This is in line with recently published research, where higher scores on motivational outcomes in the DGBL intervention can be a decisive factor for implementing/adopting DGBL, even if similar cognitive learning outcomes are achieved (All, Castellar, & Van Looy, 2015). DGBL effectiveness studies can – depending on the aim of the game, such as practicing math at home – be conducted in a context where the researcher does not have much control. Hence, other control mechanisms specific to the game environment need to be implemented, such as gathering log data on gameplay. Even more than in general effectiveness evaluation research is the addition of a follow-up study, in order to control for a short-term novelty effect.

An area of tension that clearly remains is the issue of internal versus external validity of studies. It is clear from the results that DGBL effectiveness research cannot be conducted using the Fisher tradition (Fisher, 1934, 1935), because the researcher cannot have complete control over the experiment. This is caused by the complexity of the environments where games are typically implemented, such as implementation of the interventions in natural collectives (e.g., existing class groups). Additionally, different unobserved variables can influence the outcome results and implementation in different contexts (e.g., games targeted to play at home, field experiments). Results of this study show that a proper balance between internal and external validity is best achieved by improving both aspects. Internal validity can be increased by reducing the influence of confounding variable during implementation of the intervention(s) and by adjusting analysis for potential sources of variability other than the experimental variance. External validity can be maximized by ensuring similarity between elements present in the real world implementation environment and the implementation for the effectiveness assessment, such as implementation in a context in which the game is intended to be used, implementation in natural collectives such as existing class groups (i.e., randomization on a classroom level), the presence of a familiar teacher in a classroom, the provision of procedural help, et cetera.

Several suggested best practices, however, need to be further evaluated with regard to DGBL. For instance, keeping time exposed to intervention equal in the experimental and control group might seem like a general good practice. However, in DGBL, this idea is not as straightforward because playtime does not equal time that is spent on the learning content. Moreover, an important desired outcome of implementing DGBL is cost-efficiency, of which improved time management (i.e., a reduction in time spent to learn a certain content matter) is an important indicator (All et al., 2015). When keeping time spent on the intervention equal, this prevents researchers from testing if DGBL scores better on time management. Another element, on which similarity between conditions might not be straightforward, is the reduction of support to procedural help and thus leaving out any help providing substantive information. This might jeopardize ecological validity for the control group when this is a traditional class, for instance, where asking questions related to the learning content is a common practice. Also, by providing the control group with a game training session in order to safeguard similarity between conditions, the psychological impact (e.g., in relation to motivation) of preparing participants for a game they do not get to play is ignored, which in turn, can be a confounding factor.

In conclusion, it is clear that a more standardized approach is not only possible but required in order to be able to improve rigorousness of DGBL effectiveness research and define guidelines. For instance, in order to be able to conduct a power analysis to determine which sample size one needs for his/her research, assumptions on the magnitude of the effect of the DGBL intervention need to be made. However, this assumption on the magnitude of the effect is a difficult exercise for researchers to make, as in one study the control group did not receive any educational intervention and in another study the control group got instructed using a 'classical' approach. Elements like this limit the field to grow from evaluation into research, which requires more rigorous standards of validity and reliability (Hutchinson, 1999). With this study we hope to have taken an important step in this direction.

5. Future research & limitations

Future research should focus on the development of a research protocol that supports DGBL effectiveness research. The present study took a first step in this direction by defining best practices. An interesting venue for further research would be to compare results of different study designs in published and unpublished (in order to account for publication bias) DGBL effectiveness research against our proposed best practices.

A limitation of the present study is that we cannot make claims about the representativeness of these results due to the limited number of participants and potential geographical bias. We have aimed to achieve acceptable coverage by including both national and international experts and by gathering data until saturation was achieved. An interesting approach to improve representativeness, would be conducting a survey study among a larger sample of expert in order to quantitatively validate our results.

The present study focuses on a research design for a summative evaluation and aims to determine if a DGBL intervention is successful or not. It does not provide any indication on why an intervention worked or did not work. The addition of formative evaluation research for determining what makes a DGBL intervention effective is recommended. The present study has also not taken into account the growing area of stealth assessment (i.e., unobtrusively gathering in game information as indicators for learning and motivation). Future research should definitely focus on combining results of effectiveness' studies using an experimental design and the results from stealth assessment.

Lastly, the best practices brought forward are based on the experience and expertise of academic researchers, suggesting an ideal design. In practice, however, some issues may occur which prohibit the implementation of the suggested recommendations. Hence, the authors aim to validate these recommendations by implementing them in DGBL effectiveness studies in different contexts (e.g., school, workplace, health context). This way, the recommendations can be optimized in order to develop a standardized procedure for assessing the effectiveness of DGBL that can be flexibly used across different sectors.

Acknowledgments

The current research paper is part of a Ph.D. project (grant number 121099) which is funded by the IWT, the Flemish government agency for Innovation by Science and Technology (IWT). The sponsor's role is limited to an evaluation after the second year of the Ph.D. project. Consequently, the sponsor is not involved in the study design, data collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.compedu.2015.10.007.

References

All, A., Castellar, E. P. N., & Van Looy, J. (2014). Measuring effectiveness in digital game-based learning: a methodological review. *International Journal of Serious Games, 1*(1), 3–20.

All, A., Castellar, E. P. N., & Van Looy, J. (2015). Towards a conceptual framework for assessing the effectiveness of digital game-based learning. *Computers & Education, 88*, 29–37.

Backlund, P., & Hendrix, M. (2013). Educational games-are they worth the effort? a literature survey of the effectiveness of serious games. In *Games and virtual worlds for serious applications (VS-GAMES), 2013 5th international conference on* (pp. 1–8). IEEE.

Baker, S. E., & Edwards, R. (2012). *How many qualitative interviews is enough*.

Baranowski, T., Buday, R., Thompson, D. I., & Baranowski, J. (2008). Playing for real: video games and stories for health-related behavior change. *American Journal of Preventive Medicine, 34*(1), 74.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: an overview. *Advances in Human-Computer Interaction, 2013*, 1.

Cagiltay, N. E., Ozcelik, E., & Ozcelik, N. S. (2015). The effect of competition on learning in games. *Computers & Education, 87*, 35–41.

Calder, J. (2013). *Programme evaluation and quality: A comprehensive guide to setting up an evaluation system*. Routledge.

Castellar, E. N., All, A., de Marez, L., & Van Looy, J. (2015). Cognitive abilities, digital games and arithmetic performance enhancement: a study comparing the effects of a math game and paper exercises. *Computers & Education, 85*, 123–133.

Clark. (2007). Learning from serious games? arguments, evidence, and research suggestions. *Educational Technology, 47*(3), 56–59.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2015). Digital games, design, and learning a systematic review and meta-analysis. *Review of Educational Research* (in press), 0034654315582065.

Connolly. (2014). *Psychology, pedagogy, and assessment in serious games*. IGI Global.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*(2), 661–686.

Corsi, T. M., Boyson, S., Verbraeck, A., van Houten, S.-P., Han, C., & MacDonald, J. R. (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., & 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.

Cruz, E. M. C., Cruz, J. A. V., Ruiz, J. G. R., David, L., & Hernández, H. (2015). Video games in teaching-learning processes: a brief review. *International Journal of Secondary Education, 2*(6), 102.

De Freitas, S. (2006). *Learning in immersive worlds*. London: Joint Information Systems Committee.

Fisher, R. A. (1934). *Statistical methods for research workers*.

Fisher, R. A. (1935). *The design of experiments*.

Flick, U. (2009). *An introduction to qualitative research*. Sage.

Flick, U. (2011). *Introducing research methodology: A beginner's guide to doing a research project*. Sage.

Giessen, H. W. (2015). Serious games effects: an overview. *Procedia-Social and Behavioral Sciences, 174*, 2240–2244.

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

Hays, R. T. (2005). *The effectiveness of instructional games: A literature review and discussion*. Technical Report 2005-004.

Higgins, J. P., Green, S., & Collaboration, C. (2008). *Cochrane handbook for systematic reviews of interventions* (Vol. 5). Wiley Online Librar.

Hutchinson, L. (1999). Evaluating and researching the effectiveness of educational interventions. *BMJ: British Medical Journal, 318*(7193), 1267.

Juul, J. (2003). The game, the player, the world: looking for a heart of gameness. In *Level up: Digital games research conference proceedings* (Vol. 120, p. 121). Utrecht, Holland: Utrecht Univ..

Kirriemuir, J., & McFarlane, A. (2004). *Literature review in games and learning*. Bristol, UK: FutureLab.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*(2), 311.

Kretschmann, R. (2012). Digital sport-management games and their contribution to prospective sport-managers' competence development. *Advances in Physical Education, 2*(4), 179–186.

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., Ruijven, T., et al. (2014). The research and evaluation of serious games: toward a comprehensive methodology. *British Journal of Educational Technology, 45*(3), 502–527.

Neys, J., Van Looy, J., De grove, F., & Jansz, J. (2012). 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*.

Nussbaum, M., & Beserra, V. d. S. (2014). Educational videogame design. In *Advanced learning technologies (ICALT), 2014 IEEE 14th international conference on* (pp. 2–3). IEEE.

Prensky, M. (2001). *Digital game-based learning*. New York, NY: McGraw-Hill.

Randel, J. M., Morris, B. A., Wetzel, C. D., & Whitehill, B. V. (1992). The effectiveness of games for educational purposes: a review of recent research. *Simulation & Gaming, 23*(3), 261–276.

Serrano-Laguna, Á., Torrente, J., Manero, B., Blanco, Á. d., Borro-Escribano, B., Martínez-Ortiza, I., et al. (2013). Learning analytics and educational games: lessons learned from practical experience. In *Games and learning alliance conference, Paris*.

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

Stewart, J., Bleumers, L., Van Looy, J., Mariën, I., All, A., Schurmans, D., et al. (2013). In *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.

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Sage Publications, Inc.

Suddaby, R. (2006). From the editors: what grounded theory is not. *Academy of Management Journal, 49*(4), 633–642.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation, 27*(2), 237–246.

Van Engelenburg, G. (1999). Statistical analysis for the Solomon four-group design. *Research Report*, 99–106.

Yip, F. W. M., & Kwan, A. C. M. (2006). Online vocabulary games as a tool for teaching and learning English vocabulary. *Educational Media International, 43*(3), 233–249.