Journal of Universal Computer Science, vol. 22, no. 7 (2016), 992-1021
submitted:11/12/15, accepted: 26/5/16, appeared: 1/7/16 © J.UCS

How to Evaluate Educational Games: a Systematic Literature Review

Giani Petri
(Graduate Program in Computer Science, Federal University of Santa Catarina
Florianópolis/SC, Brazil
giani.petri@posgrad.ufsc.br)

Christiane Gresse von Wangenheim
(Graduate Program in Computer Science, Federal University of Santa Catarina
Florianópolis/SC, Brazil
c.wangenheim@ufsc.br)

Abstract: Educational games have been used as an innovative instructional strategy in order to achieve learning more effectively. However, it is essential to systematically evaluate such games in order to obtain sound evidence on their impact. Thus, the objective of this article is to present the state of the art on how to systematically evaluate educational games. Therefore, we performed a systematic literature review with an initial sample of 21,291 articles from which 11 relevant articles have been identified, describing 7 approaches to systematically evaluate educational games. Based on these studies we analyze how the approaches are defined (quality factors, theoretical constructs), operationalized (research designs, data collection instruments, data analysis methods), how they have been developed (development methodology) and evaluated (evaluated aspects, number of applications & data points and data analysis methods). As a result, we can confirm that exist few approaches to systematically evaluate educational games. The majority of the approaches are developed in an ad-hoc manner, not providing an explicit definition of the study, its execution and data analysis. We also observed that among the few encountered approaches no clear pattern emerges on which quality factors to evaluate. This shows that there exists a need for research on the definition and operationalization of educational game evaluations in order to obtain more valid and uniform results.

Keywords: educational games, serious games, games evaluation, systematic literature review, state of the art
Categories: L.0.0, L.5.1

1 Introduction

Educational games (or serious games) are specifically designed to teach people about a certain subject, expand concepts, reinforce development, or assist them in drilling or learning a skill or seeking a change of attitude as they play [Dempsey, 96a]. In recent years, educational games have been used as an innovative instructional strategy in order to achieve more effectively learning on higher levels [Connolly, 12; Calderón, 15] in diverse knowledge areas, such as mathematics, language, business, health, computing, nutrition, firefighting, entertainment among others [Connolly, 12; Backlund, 13; Calderón, 15, Battistella, 16].

Educational games are believed to result in a wide range of benefits, like increasing the learning effectiveness, interest and motivation as well as a reduction of training time and instructor load [Garris, 02; Prensky, 07; Wangenheim, 09a, Wouters, 13; Hamari, 16]. They are expected to be a fun and safe environment, where students can try alternatives and see the consequences, learning from their own mistakes and practical experiences [Pfahl, 01]. Thus, they are supposed to be an effective and efficient instructional strategy for teaching and learning. Yet these claims are questionable or at least not rigorously established [Hays, 05; Akili, 06; Wangenheim, 09a; All, 16]. But in order to use them effectively it is essential to systematically evaluate such games in order to obtain sound evidence on their impact.

In this context, a number of literature reviews of educational games have been carried out to elicit empirical evidence of the games' impact on student learning [Connolly, 12; Backlund, 13; Calderón, 15; Wangenheim, 09a; Wangenheim, 09b; Caulfield, 11; Gibson, 13]. Their findings show that most evaluations of educational games are performed in an ad-hoc manner in terms of research design, measurement and data collection & analysis lacking scientific rigor. However, few studies present a review of the existing well-defined approaches to systematically evaluate educational games, with the exception of [Connolly, 09]. In this article, the authors present a literature review on evaluation approaches for games-based learning [Connolly, 09] in which they identify 6 approaches. However, no details on the encountered approaches are presented and as the review has been conducted in 2008, new approaches may be available nowadays.

Thus, in order to elicit the state of the art of how to systematically evaluate educational games, we conducted a systematic literature review. The main contribution of this paper is the analysis and summary of how the approaches are defined (quality factors, theoretical constructs), operationalized (research designs, data collection instruments, data analysis methods), developed (development methodology) and evaluated (evaluated aspects, number of applications, data points and data analysis methods), expecting to find theoretical and operational patterns. The results of this review may assist game designers and/or instructors to systematically evaluate educational games in order to obtain feedback and to identify improvement opportunities as well as to guide the application of games in educational practice.

2 Background

2.1 Educational games

An educational game is an instructional strategy that involves competition and is organized by rules and restrictions to achieve a certain educational goal [Dempsey, 96b]. They are specifically designed to teach specific concepts or to strengthen competencies [Abt, 02]. The use of games as instructional strategy is also known as game-based learning [Prensky, 07].

Educational games are characterized by various elements, such as goals, rules, restrictions, interaction, challenge, competition, rewards and feedback [Prensky, 07; Wangenheim, 12]. Intrinsic characteristics of games, such as competition stimulating the will to win, help students to stay focused on the learning activity [Prensky, 07].

There exist a broad scope of games including digital and non-digital ones [Connolly, 12; Backlund, 13; Calderón, 15; Boyle, 16; Battistella, 16]. Digital games are developed for use in smartphones, computers, tablets, etc. [Mitamura, 12], whereas non-digital games exploring the use of resources such as boards, cards, pencils and papers [Connolly, 07]. Table 1 presents a classification in terms of platform of the different types of educational games, based on [Caulfield, 11; Connolly, 12; Battistella, 15].

Category			**Definition**
Digital game			Electronic game that involves human interaction with a user interface to generate visual feedback on an electronic device.
	PC game	stand-alone	Game played on a general-purpose personal computer.
		online	Game played on some form of computer network (Internet), using a personal computer.
	Console game		Game played on a specialized electronic device that connects to a common television set or composite video monitor.
	Mobile game		Game played on a mobile device, such as, phone, tablet media player, etc.
Non-digital game			Game that is not played on an electronic device.
	Board game		Game that involves counters or pieces moved or placed on a pre-marked surface or "board", according to a set of rules.
	Card game		Game using playing cards as the primary device with which the game is played.
	Paper & pencil game		Game that can be played solely with paper and pencil.
	Prop game		Game that is played using props (portable objects).

Table 1: Game platforms [Battistella, 15].

Besides of classifying in terms of platforms, games also cover a broad spectrum of genres such as action, adventure, strategy, simulation, puzzle, quiz, role-playing (RPG), among others [Herz, 97]. In addition, another characteristic is the interaction mode between the game and players, the game interaction mode is typically classified in single-player, multi-players or multi-groups [Battistella, 15].

In recent years, educational games also have been used to teach competences in higher education context in diverse knowledge areas, such as health & wellness, culture, social skills, professional learning & training, among others [Rodriguez-Cerezo, 14; Calderón, 15; Soflano, 15]. The type of game that attracts more interest is digital/computer games, principally PC games, yet, with a considerable trend also to non-digital games (paper & pencil, board games, etc.) [Calderón, 15, Battistella, 16]. In computing education, for example, simulation games, that allow to practice competencies in a realistic environment while keeping students engaged, are predominant. Most games aim at learning objectives on lower cognitive levels, often being used to review and reinforce knowledge taught beforehand using different instructional strategies and/or the simulation of real-life situations teaching competencies on the application level [Wangenheim, 09a; Battistella, 16].

2.2 Evaluation of Instructional Strategies

The evaluation of an instructional strategy aims at measuring the level of its success, evaluating whether the target audience achieved the defined objectives. Evaluation should cover both the student learning, as well as the quality of elements, materials and resources that compose the instructional strategy [Branch, 10].

Instructional strategies can be evaluated through analytical or empirical methods. An analytical evaluation consists in an inspection performed by a group of experts [Preece, 02]. This type of evaluation is characterized by the non-involvement of end users and aims at identify potential issues. On the other hand, empirical studies involve end users collecting data while applying the instructional strategy. This is typically done in form of surveys, case studies or experiments [Wohlin, 12].

Depending on the objective of the evaluation, models, methods, scales or frameworks can be used to conduct the research [Hevner, 10]:

- A model consist of sets of propositions or statements expressing relationships between constructs (the conceptual vocabulary of a domain).
- A method consist of sets used to perform tasks.
- A framework is used as real or conceptual guide to serve as support.
- A scale is an effective instrument to measure variables.

A systematic evaluation following the process of an empirical study involves several phases, such as scoping, planning, operation, analysis & interpretation, and presentation & package. In the scoping phase, the evaluation objective and goals are defined. This includes the explicit specification of the quality factors to be evaluated such as learning, engagement, motivation, etc. In the planning phase, an appropriate research design is defined identifying also the level of evaluation based, for example, on the four-level model for evaluation [Kirkpatrick, 06], as shown in Table 2.

Level		**Evaluation description and characteristics**	**Examples of evaluation methods and instruments**
1	Reaction	Evaluates how the participants felt about the training or learning experience.	Feedback forms; verbal reactions; post-training questionnaires
2	Learning	Evaluates the increase in knowledge or skills.	Reviews and tests before and after training; interview and observations
3	Behavior	Evaluates the degree to which new learning acquired actually transfers to the job performance.	Observations and interviews over time to assess changes
4	Results	Evaluation of the effect on the business environment by the learner.	Observation and measurement over time; interviews with participants, their managers and customer groups

Table 2: Four-level model for evaluation [Kirkpatrick, 06].

Common study types and research designs used in evaluations in education contexts are summarized in Table 3.

Evaluation level [Kirkpatrick, 06]	**Study type**	**Design**	**Representation** **X=Treatment** **O=Measurement** **R=Random assignment**
1 – Reaction: Evaluates how the participants felt about the training or learning experience.	Non-experimental	One-shot post-test only	X O
2 – Learning: Evaluates the increase in knowledge or skills.	Non-experimental	One-shot pre-test/post-test	O X O
	Quasi - experimental	Static group comparison group	X O O
		Static group pre-test – post-test	O X O O O
		Times Series	O O X O O
	Experimental	Randomizes post-test only	R X O R O
		Randomized pre-test/post-test	R O X O R O O
		Randomized pre-test/post-test control group	R O X1 O R O X2 O

Table 3: Common types of research design [Wangenheim, 09a; Shadish, 02].

In order to achieve the evaluation goal(s), measurement has to take place [Fenton, 98; Wohlin, 12]. Therefore, measures and data collection instruments have to be defined in a way that allows to trace the evaluation goal to the data to be collected and also provides a framework for analyzing and interpreting the data with respect to the goals. The operation phase includes the preparation and execution of the study by applying the treatment (the educational game and optionally other instructional strategies for comparison) and collecting data as defined. During the analysis & interpretation phase, the collected data is analyzed with respect to the evaluation goal(s). Depending on the nature of the collected data, this may be done by using qualitative and/or quantitative analysis methods ranging from descriptive statistics to inferential statistics as summarized in Figure 1 [Wohlin, 12; Freedman, 07].

The analyzed data is interpreted, answering the analysis questions and, consequently, obtaining the evaluation goal.

3 Methodology

In order to elicit the state of the art on how to systematically evaluate educational games, we performed a Systematic Literature Review (SLR) following the procedure defined by [Kitchenham, 10]. A SLR uses systematic and explicit methods to identify, evaluate and interpret all relevant studies for clearly defined research questions [Clark, 03; Kitchenham, 10]. Figure 2 illustrates the steps of the adopted SLR process.

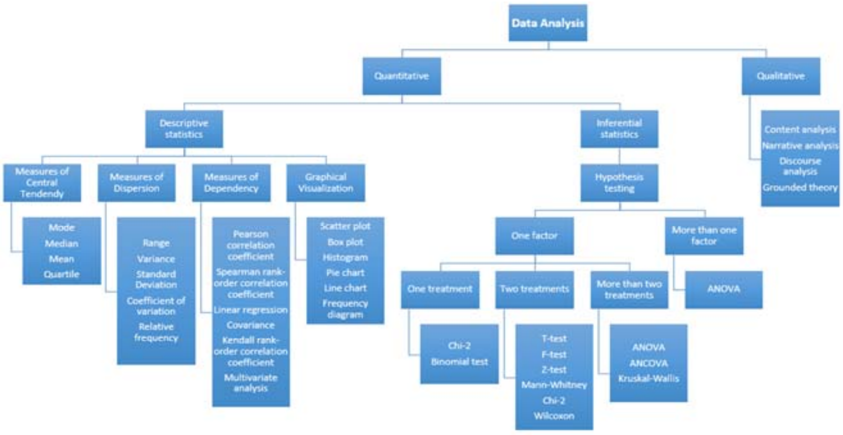


Figure 1: Data analysis methods

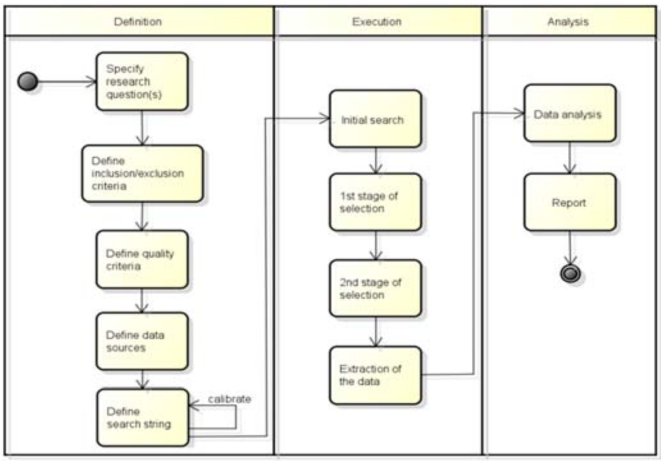


Figure 2: Activity diagram of SLR process (adapted from [Kitchenham, 10])

As shown in Figure 2, the SLR process is divided into three phases: definition, execution and analysis. In the definition phase (Section 4), research objectives are identified and a systematic review protocol is defined. The protocol specifies the central research questions and the procedures that will be used to conduct the review, including the definition of inclusion/exclusion criteria, quality criteria, data sources, and search string. The execution phase (Section 5) consists of the search and identification of relevant studies, and their selection in accordance to the inclusion/exclusion and quality criteria established in the protocol. Once identified, data related to the research question(s) is extracted from the relevant studies, analyzed and synthesized during the analysis phase (Sections 6 and 7).

4 Definition of the Systematic Literature Review

This research aims at the elicitation of the state of the art on how to systematically evaluate educational games. In accordance to this objective, we performed a SLR, following the steps defined in Section 3, focusing on the following research questions.

Research Questions
RQ1: Which models, methods, scales, or frameworks (approaches) exist to systematically evaluate educational games?
RQ2: Which quality and/or sub-quality factors are evaluated?
RQ3: How data collection and analysis is operationalized?
RQ4: How these approaches have been developed?
RQ5: How these approaches have been evaluated?

Inclusion/Exclusion Criteria. In accordance to our research objective/questions, criteria for selecting only relevant studies were defined. We included only articles that presented a well-defined approach to systematically evaluate educational games for teaching any knowledge area. We focused only on articles of empirical studies/evaluations, written in English (or in Portuguese with an abstract in English), available via digital libraries published during the last 20 years (between January 1995 and October 2015).
On the other hand, we excluded:

- Any study not related to well-defined approach;
- Any study not related to an empirical study/evaluation;
- Articles that present the evaluation of an educational game, but do not use a well-defined approach.

Quality criteria. In addition to our inclusion/exclusion criteria, we also superficially assessed the quality of the reported studies, considering only articles that provide substantial information on the evaluation approach.

Data Sources and Search String. Data sources have been chosen based on their relevance in the computing domain, including: ACM Digital Library, IEEE Xplore, Springer Link, Science Direct and Wiley Online Library. In addition, we also searched via Google Scholar, in order to also consider articles published outside the computing domain, but which may provide a relevant contribution.

In accordance to our research objective, we defined the search string by identifying core concepts such as model, educational games, and evaluation including also synonyms as indicated in Table 4.

Core Concepts	Synonyms
model	method, framework, scale
educational games	serious games, game-based learning
evaluation	assessment

Table 4: Keywords

Using these keywords, the search string has been calibrated and adapted in conformance with the specific syntax of each of the data sources as presented in Table 5.

Data source	Search String
ACM Digital Library	(model OR method OR framework OR scale) AND ("educational games" OR "serious games" OR "game-based learning") AND (evaluation OR assessment) for: ((model OR method OR framework OR scale) AND ("educational games" OR "serious games" OR "game-based learning") AND (evaluation OR assessment)) Published since January 1995
IEEE Xplore	((model OR method OR framework OR scale) AND ("educational games" OR "serious games" OR "game-based learning") AND (evaluation OR assessment) IN metadata) AND (pyr >= 1995 AND pyr <= 2015)
Springer Link	'(model OR method OR framework OR scale) AND ("educational games" OR "serious games" OR "game-based learning") AND (evaluation OR assessment)' published between 1995 - 2015 within Article
Science Direct	pub-date > 1994 and ((model OR method OR framework OR scale) AND ("educational games" OR "serious games" OR "game-based learning") AND (evaluation OR assessment))
Wiley Online Library	(model OR method OR framework OR scale) AND ("educational games" OR "serious games" OR "game-based learning") AND (evaluation OR assessment) in All Fields between years 1995 and 2015 Publication type: Journals
Google Scholar	(model OR method OR framework OR scale) AND ("educational games" OR "serious games" OR "game-based learning") AND (evaluation OR assessment) Custom range: 1995-2015

Table 5: Search Strings

5 Execution of the Review

The SLR was conducted in October and November 2015 by the first author, a Computer Science Ph.D. candidate, and was reviewed by a senior researcher (second author). In the initial search, we found a total of 21,291 articles. Table 6 summarizes the returned results per data source. From Google Scholar we selected only the 1,000 most relevant results (100 first pages), from ACM Digital Library and Science Direct the 1,000 most relevant results, observing a lack of relevancy after these quantities.

From IEEExplore, SpringerLink and Wiley online library all returned articles were analyzed. As result, a total of 4,738 articles were analyzed during the first stage.

	Google Scholar	ACM	IEEExplore	Springer Link	Science Direct	Wiley	Total
Initial Search	17,000	1,314	138	791	1,239	809	21,291
Total analyzed during 1st stage	1,000	1,000	138	791	1,000	809	4,738
Selected after 1st stage	68	8	14	6	15	8	119
Selected after 2nd stage	5	0	2	0	3	1	11

Table 6: Search results

During the first stage, the search results were quickly analyzed based on their title and short summary. The abstract was read only in case the title did not provide evidence of any exclusion criteria. Irrelevant and duplicate papers were removed. This stage left us with 119 potentially relevant articles. Then, we performed a second stage of selection. In this stage, we analyzed the full abstract of the articles and quickly scoped the article for information on the evaluation approach. In this stage, we excluded articles focusing only on heuristic/usability evaluation of educational games [Gunter, 08; Omar, 08; Omar, 10] and/or non-educational video games [Pinelle, 08; Sweetser, 12]. As a result, 11 articles (describing a total of 7 approaches) were identified as primary studies. The complete list of relevant articles is available in Appendix 1 and 2.

6 Data Extraction

In accordance with the defined research questions, we systematically extracted information in a spreadsheet from each article selected for analysis. Table 7 shows the data items that were extracted.

The articles were read thoroughly and data was extracted by the first author and reviewed by the second author. Data extraction was hindered in several cases by the way in which the studies were reported. Most papers lack sufficient detail about the definition, development and validation of the evaluation approach. In some cases, more than one evaluation approach was reported in one article, or the same approach was reported by more than one article. In these cases, we extracted the information on each of the approaches separately. A complete overview of the extracted data is available in Appendix 1 and 2.

Research question	**Data Item**	**Description**
RQ1: Which models, methods, scales, or frameworks (approaches) exist to systematically evaluate educational games?	Reference	Reference of the study.
	Name	Acronym or name of the approach.
	Instructional strategy	The instructional strategy focused by approach.
	Overview	A graphic overview of the approaches.
RQ2: Which quality and/or sub-quality factors are evaluated?	Quality (sub-) factor(s)	Quality (sub-) factor(s) that are evaluated.
	Theoretical basis	The theoretical construct(s) used to define the quality factors that are evaluated.
RQ3: How data collection and analysis is operationalized?	Study type	Study type classified based on Table 3 following common research designs used in education contexts.
	Data collection instrument(s)	Instrument(s) used for data collection, such as questionnaires, interviews, or observations.
	Response format	Type of measurement scales used for data collection.
	Data analysis method(s)	Method(s) used for data analysis based on the classification presented in Figure 1.
RQ4: How these approaches have been developed?	Development methodology	Methodology used to develop the approach.
RQ5: How these approaches have been evaluated?	Evaluated factors	Factors used to evaluate the approach.
	Number of applications	Number of studies applying the approach.
	Data points	Number of data points collected during the applications used to evaluate the approach.
	Data analysis method(s)	Method(s) used for data analysis to evaluate the approach.
	Findings	Brief description of the principal results of the study.

Table 7: Data items extracted

7 Data Analysis

In total, we identified 11 articles describing 7 approaches to evaluate educational games. Although we considered the last 20 years (1995-2015) in our review, we only encountered relevant publications in the last 10 years, after 2006, as shown in Figure 3. This shows that the interest in approaches to systematically evaluate educational games has been growing in the last years.

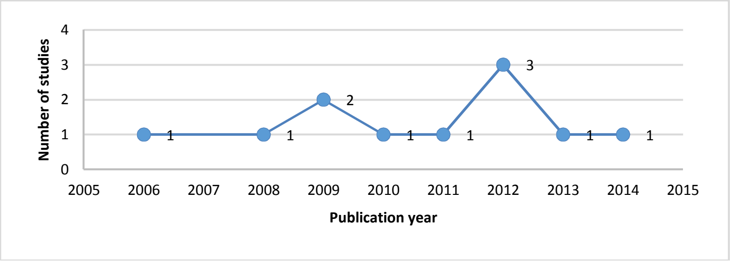


Figure 3: Distribution of studies per year of publication.

In order to present our findings, we analyze each of the research questions separately.

RQ1: Which models, methods, scales, or frameworks (approaches) exist to systematically evaluate educational games?

Analyzing the selected studies, we identified 7 different approaches to systematically evaluate educational games. Three approaches present a framework [Connolly, 09; Freitas, 06; Carvalho, 12], two approaches present a scale [Fu, 09; Ak, 12], one approach presents a method [Mayer, 12] and another approach presents a model [Savi, 11]. We present a brief description of each one approach.

The Evaluation Framework for Effective Games-based Learning (GBL) [Connolly, 09] is a framework for GBL based on key measurements identified in the literature. The purpose of the framework is to identify what can potentially be evaluated in a GBL application. The approach proposes the evaluation of GBL with respect to learner performance, learner/academic motivation, learner/academic perceptions, learner/academic preferences, the GBL environment itself and the collaboration between players. The framework can be customized to particular requirements depending on particular analytical measurement is needs.

Another approach is the four-dimensional framework [Freitas, 06]. This framework helps tutors to evaluate the potential of using games- and simulation-based learning in their practice. The framework allows practitioners to be more critical about how they embed games and simulations into their lesson plans. It allows researchers and evaluators to develop metrics for supporting effective analysis of existing educational games and simulations and allows educational designers to consider a more user-based and specialized set of educationally specific factors. The four dimensions evaluated by the framework are: context, learner or learner group, internal representation world, and process of learning.

Carvalho (2012) presents an evaluation framework that assesses the efficiency of GBL focusing on engineering education [Carvalho, 12]. Covering the two first levels of Kirkpatrick's evaluation model (reaction and learning) [Kirkpatrick, 06], the framework is divided in three stages: alpha-testing, beta-testing and gamma-testing each with clear objectives, predefined protocols and data collection tools. The framework assesses the games' efficiency in terms of game play, game story, mechanisms, usability, knowledge, motivation, and satisfaction.

Fu et al. (2009) present EGameFlow [Fu, 09], a scale that assesses user enjoyment of e-learning games to help developers to understand strengths and weaknesses from the students' perception in accordance to evaluation level 1 (reaction) [Kirkpatrick, 06]. It evaluates the game's quality with respect to eight factors: immersion, social interaction, challenge, goal clarity, feedback, concentration, control, and knowledge improvement.

Another scale was proposed by [Ak, 12]. This scale aims at the selection of good educational computer games. The scale is intended to measure the quality of games before applying it in class. Game quality is measured in terms of enjoyment and learning.

A comprehensive methodology for the research and evaluation of serious games was proposed by [Mayer, 12]. This generic evaluation methodology for serious gaming, consists of a framework, conceptual models, research designs, evaluation constructs and scales, and data collection techniques. The methodology assesses serious games in three different moments (pre-game, in-game, and post-game) in terms of previous experiences/skills, game performance, game play, game experience, player satisfaction, and learning.

Another methodology is also provided based on the model MEEGA (Model for the Evaluation of Educational Games) [Savi, 11] that is specifically developed for the evaluation of educational games. The model focuses on evaluation level 1 (reaction) [Kirkpatrick, 06], capturing the reaction of students after they played the game by applying a standardized questionnaire. MEEGA measures three quality dimensions of educational games: motivation, user experience, and learning. The model is also accompanied by a process on how to apply the evaluation model in practice.

RQ2: Which quality and/or sub-quality factors are evaluated?

In order to answer this question, we analyzed the quality and/or sub-quality factors evaluated by the identified approaches. In summary, we identified 52 different quality and/or sub-quality factors that have been used by the approaches to evaluate educational games. All approaches use more than one quality factor to evaluate the games. The evaluated quality factors are shown in Figure 4 indicating their frequencies by the size of the factor (and the associated number).

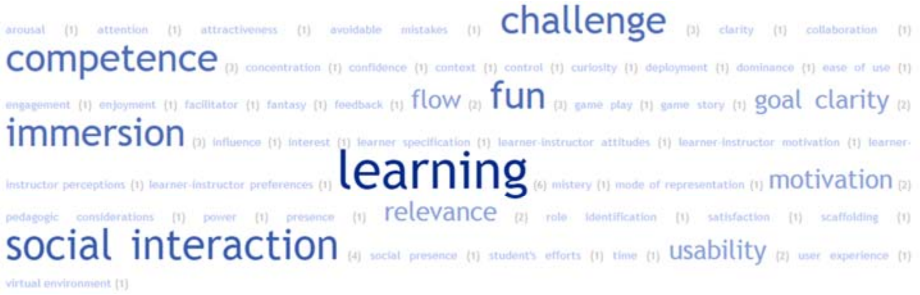


Figure 4: Frequencies of quality and/or sub-quality factors used.

As shown in Figure 4, we can observe a wide diversity of quality factors evaluated not allowing the identification of any clear pattern. Another issue is that only very few studies systematically derivate the metrics by decomposing the quality factors based on theoretical constructs, as e.g., MEEGA [Savi, 11], using the ARCS (Attention, Relevance, Competence, Satisfaction) model [Keller, 87] to decompose the motivation concept. Therefore, we present a list of all quality and/or sub-quality factors identified in the selected studies (instead of a hierarchical decomposition) in Table 8.

Quality/sub-quality factor	Quantity of approaches considering this factor	Description
Learning	6	The evaluation of this factor refers to the improvement of competence. Only one study [Savi, 11] evaluates the learning effect with regard to a systematic definition of learning levels based on Bloom's taxonomy [Bloom, 56]. [Connolly, 09] defines learning as learner performance, that an improvement in the performance of the learner as a result of the intervention.
Social Interaction	4	Social interaction refers to the creation of a feeling of shared environment and being connected with others in activities of cooperation or competition [Fu, 09].
Challenge	3	Challenge means that a game needs to be sufficiently challenging with respect to the player's competency level. The increase of difficulty should occur at an appropriate pace accompanying the learning curve. New obstacles and situations should be presented throughout the game to minimize fatigue and to keep the students interested [Sweetser, 05].
Competence	3	Players need to realize that their competencies are at a level where it is possible to overcome the challenges of the game. As the difficulty increase, challenges should require the player to develop their competencies to advance in the game and have fun [Sweetser, 05; Poels, 07; Takatalo, 10].
Immersion	3	Immersion allows the player to have an experience of deep involvement within the game, creating a challenge with real-world focus, so that s/he forgets about the outside world during gameplay [Fu, 09].
Fun	3	Fun refers to students' feeling of pleasure, happiness, relaxing and distraction [Poels, 07].
Relevance	2	Relevance refers to the students need to realize that the educational proposal is consistent with their goals and that they can link content with their professional or academic future [Keller, 87].
Goal Clarity	2	Games need to have clearly defined goals with manageable rules [Fu, 09].
Usability	2	[Carvalho, 12] defines usability in terms of awareness of progress, consistence of interface (colors, fonts), controls and visual feedback. [Connolly, 09] refers to Usability in terms of task completion times, average task completion times, the ease of the task, the number of errors made while performing a task and the ranking of the tasks by the learners.
Motivation	2	Concept used in a general way without any further decomposition [Carvalho, 12; Mayer, 12].
Flow	2	Games should be designed to generate positive affects in players facilitating flow experience [Kiili, 05], where flow is

		a state in which complete absorption or engagement is realized [Csikszentmihalyi, 91; Ak, 12].
Satisfaction	1	Satisfaction means that the students must feel that the dedicated effort results in learning [Keller, 87].
Attention	1	Attention refers to students' cognitive responses to instructional stimuli. It is desirable to obtain and maintain a satisfactory level of attention of students during a learning period [Keller, 87].
Confidence	1	Confidence means to enable students to make progress in the study of educational content through their effort and ability (e.g., through exercises with increasing level of difficulty) [Keller, 87].
Concentration	1	Games need to screen out distraction and make concentration possible [Fu, 09].
Feedback	1	Games need to provide clear information on how the participants are doing [Fu, 09].
Arousal	1	[Sherry, 06] state that arousal is an emotion that results from fast actions and high quality graphics [Ak, 12].
Collaboration	1	Refers to the achievement of learning outcomes or particular goals by log files monitoring interaction, mapping team aspects to learner comments, measuring the regularity and level of collaboration and learner group reflection essays [Connolly, 09].
Control	1	Games need to allow the player to have a sense of control over the interactions that should be easy to learn and allow them to explore the game freely and at their own pace [Poels, 07; Takatalo, 10].
Game Play	1	Refers to all player experiences during its interaction with the game [Carvalho, 12].
Game story	1	Game story refers to how much the storyline is linear, clear, and interesting [Carvalho, 12].
Curiosity	1	Games need to evoke curiosity, e.g. through mysteries [Garris, 02] [Ak, 12].
Deployment	1	Deployment is intended to encompass the most effective method of incorporation of the GBL application into the educational context and can also mean the preference of different gaming conditions [Connolly, 09].
Enjoyment	1	Refers to feelings of pleasure, appreciation when playing the game [Carvalho, 12].
Fantasy	1	Fantasy is the ability to do things in the games that people are not able to do in real life such as flying, driving race cars etc. [Sherry, 06] [Garris, 02] [Ak, 12].
Interest	1	Interest refers to how much the game is interesting and attractive for the students' learning [Carvalho, 12].
Learner/instructor attitudes	1	Refers to learner and instructor attitudes towards various elements that may alter the effectiveness of the intervention. These elements include: learner attitudes towards the taught subject, learner attitudes towards games [Connolly, 09].
Learner/instructor Motivation	1	Refers to the particular motivations of the learner for using the intervention, the learner level of interest in participating in the intervention, participation over a prolonged period of time and determining what particular motivations are the most important [Connolly, 09].
Learner/instructor Perceptions	1	Refers to perceptions associated with the learners such as their perception of the overview of time within a game or simulation, how real the game is and its correspondence with reality [Connolly, 09].
Learner/instructor	1	Refers to learner preference for media when teaching the

Preferences		material, preference of conventional or GBL training, preference and utilization of particular game features, most preferred positive and negative aspects of the game and preference for different competitive modes [Connolly, 09].
Mystery	1	[Berlyne, 60] explains that mystery is a result of the incongruity of information complexity, novelty, surprise and violation of expectations [Ak, 12].
Scaffolding	1	Scaffolding refers to the advice and resources within the environment to support the learner in completing their learning outcomes [Connolly, 09].
Social Presence	1	Level of social presence is related to the immersion and interaction in the game world [Connolly, 09].
User Experience	1	[Carvalho, 12] refers to the students' emotions and attitudes about using a particular game.
Virtual Environment	1	In terms of the actual virtual environment itself the evaluation criteria may include: background environment and characters including virtual agent expressiveness, environmental alteration, advice importance within the environment, the context of the environment in terms of real-world decision making support and general game difficulty [Connolly, 09].
Pedagogic considerations	1	Refers to the processes of learning both during the course of formal curricula based learning time and during informal learning. In particular, this dimension promotes the practitioners' reflection upon methods, theories, models and frameworks used to support learning practice [Freitas, 06].
Learner specification	1	Refers to attributes of the particular learner or learner group, this may include the age and level of the group, as well as specific components of how they learn including their learning background, styles and preferences [Freitas, 06].
Context	1	Refers to the particular context where play/learning takes place, including macro-level historical, political and economic factors as well as micro-level factors such as the availability of specific resources and tools. The tutor's own specific background and understanding as well as the availability of technical support [Freitas, 06].
Mode of representation	1	Refers to the internal representational world of the game or simulation, which in this context includes: the mode of presentation, the interactivity, the levels of immersion and fidelity used in the game or simulation [Freitas, 06].
Time	1	Metric relating to in-game scores [Mayer, 12].
Avoidable mistakes	1	Metric relating to in-game scores [Mayer, 12].
Dominance	1	Gameplay metric to measure the domain of the player about the game [Mayer, 12].
Power	1	Gameplay metric to measure the power of the player about the game [Mayer, 12].
Influence	1	Gameplay metric to measure the influence of the player on the game [Mayer, 12].
Engagement	1	The game was engaging and the game sustained their engagement [Mayer, 12].
Clarity	1	Refers clarity and easy understanding of the game [Mayer, 12].
Attractiveness	1	Affect how much the product captures the user's emotional responses [Mayer, 12].
Ease of use	1	Ease of use refers to how much the game's information are clear, organized, if the students know where they are and how to get where they want, it is user friendly and simple to use [Mayer, 12].

Student's efforts	1	Refers to the student's efforts to achieve the goals of the game to win [Mayer, 12].
Role identification	1	Level of identification of its role and activities in the game [Mayer, 12].
Facilitator	1	Refers to the help / support offered by the game [Mayer, 12].
Presence	1	Refers to the level of attention captured by the game [Mayer, 12]

Table 8: Evaluated factors/sub- factors

RQ3: How data collection and analysis is operationalized?

In order to answer this question, we analyzed how the approaches operationalize the evaluation, including research designs, data collection instruments, and data analysis methods.

Analyzing the research designs, we classified the approaches in accordance to common study types as presented in Section 2.2. Two approaches [Fu, 09; Carvalho, 12] provide an evaluation approach to be conducted in an ad-hoc manner, not providing a systematic definition of the evaluation objective. Only one approach [Savi, 11] proposes the conduction of the evaluation in form of a case study (non-experimental). The process defined by MEEGA explicitly defines the evaluation objective and provides a standardized questionnaire based on the systematically defined model to be applied after the treatment (educational game) to collect data on the learners' perception. The approach proposed by [Mayer, 12] defines a quasi-experimental design, similar to the experimental design, but without a random allocation of learners to the experimental or control group.

No information on the operationalization of the evaluation was given by [Ak, 12; Connolly, 09; Freitas, 06].

Analyzing the kind of data collection instruments, we identified that the majority collect data via questionnaires (3 approaches), but only two approaches have systematically developed and statistically evaluated [Savi, 11; Fu, 09]. [Carvalho, 12] also is used a questionnaire as data collection instrument, but not provide information about its validity. Analyzing the response format of these scales we identified that the Likert scale is the most used one (3 approaches), typically, representing the lowest and highest degree to which respondents agree with the items. In addition, an ordinal scale also is used (2 approaches) to measure specific characteristics.

Others data collection methods used include semi-structured interviews [Carvalho, 12] and tests in order to assess the knowledge of the students [Carvalho, 12].

Analyzing the data analysis methods of the selected studies, only two approaches [Fu, 09; Savi, 11] provide information about which methods are used to analyze the data collected. [Savi, 11] uses descriptive statistical methods (median/mode) and graphical visualization techniques such as histogram and frequency diagrams. [Fu, 09] uses descriptive statistics methods such as mean, standard deviation, and Pearson correlation coefficient to examine the dependency between variables. In addition, this approach also includes hypothesis testing in order to reject (or accept) a hypothesis with respect to a quality factor of the game. The t-test is used to compare two sample means, in a one factor-two treatments design and ANOVA is used to evaluate the discrepancy in the level of psychological enjoyment between subjects [Fu, 09].

RQ4: How these approaches have been developed?

Analyzing the selected studies, we identified that most of approaches (5) do not report a systematic methodology to develop the approach. In general, the approaches seem to be developed in ad-hoc manner [Carvalho, 12] or only based on theoretical constructs [Ak, 12; Mayer, 12; Connolly, 09], but not providing an explicit definition of the objective, measures or data collection instruments.

On the other hand, two approaches report a systematic methodology for their development [Savi,11; Fu, 09]. MEEGA and EGameFlow follow the Scale Development Guide [DeVellis, 03] to systematically develop a measurement instrument. In addition, MEEGA has been developed by using the GQM (Goal/Question/Metric) approach [Basili, 94] to explicitly define a measurement program for evaluating three quality dimensions of educational games: motivation, user experience and learning based on theoretical constructs.

RQ5: How these approaches have been evaluated?

In order to answer this question, we analyzed the factors used to evaluate the approaches. We identified that the most of approaches (5) do not explicitly define criteria. Typically, the approaches (4) are proposed and partially evaluated through some case or pilot studies, applying the approach to evaluate an educational game in class [Carvalho, 12; Mayer, 12; Connolly, 09; Freitas, 06]. No information with respect to its evaluation was encountered for the approach proposed by [Ak, 12].

On the other hand, two approaches present a systematic evaluation [Savi, 11; Fu, 09]. MEEGA has been evaluated in terms of its applicability, utility, validity and reliability through three case studies in two different courses on three educational games [Savi, 11]. A total of 79 data points were collected and analyzed with respect to [Devellis, 03]: intercorrelation of scale items, item-total correlation, variance, mean, and Cronbach's alpha coefficient. The model was considered easy to use, requires little interruption of classes and the measuring instrument used for data collection presented a satisfactory performance on a statistical analysis of validation.

EGameFlow has been evaluated in terms of its item analysis, reliability and validity through 4 games sessions in the same course, using different e-learning games [Fu, 09]. A total of 166 data points were collected and analyzed using the following tests: mean, standard deviation, extreme group comparison, test for homogeneity, t-test, ANOVA, Pearson's correlation, and Cronbach's alpha correlation. Statistical analyzes showed that the scale developed demonstrates high validity and reliability, which makes it an effective tool for assessing e-learning games.

8 Discussion

Considering that in the last 20 years, only 7 evaluation approaches have been encountered used in 11 studies, indicates the need for a more large-scale application of such evaluations of educational games. We can observe that there exist a few number of approaches to systematically evaluate educational games.

Analyzing the selected approaches (RQ1), we identified that most propose a framework (3 approaches) to systematically evaluate educational games [Connolly, 09, Freitas, 06; Carvalho, 12]. Typically the frameworks define a set of criteria ranging from pedagogical perspective to gaming perspective, including context, environment, learner specifications, preferences, game play, user experience, etc. [Freitas, 06; Connolly, 09; Carvalho, 12]. These criteria are used to guide and to help support instructors/tutors to evaluate educational games in a particular learning context and knowledge area [Freitas, 06]. Thus, these frameworks are considered a flexible and easy to use approach, with the ability to help practitioners to reflect upon learning processes and approaches [Freitas, 06]. However, the frameworks itself do not provide guidance on how to conduct the evaluation, data collection and analysis.

In this regard, the work presented by [Fu, 09; Ak, 12] propose scales providing effective instruments to systematically measure the quality of the games [Fu, 09]. However, only the EGameFlow scale [Fu, 09] has been evaluated analyzing its validity and reliability as an effective tool to evaluate the level of enjoyment provided by e-learning games to their users [Fu, 09]. On the other hand, no evaluation of the scale proposed by [Ak, 12] has been encountered, thus, leaving its validity and reliability questionable [Kitchenham, 95; Kimberlin, 08].

As the most comprehensive support, two methods have been encountered. [Mayer, 12] proposes a generic evaluation method for serious game. But, although the method provides comprehensive support, including a framework, conceptual models, research designs, evaluation constructs and scales, and data gathering techniques, no information on the applicability and validity of this method have been encountered. On the other hand, the MEEGA [Savi, 11] provides an evaluation method by the evaluation process based in the MEEGA model that has been systematically developed by using the GQM approach to explicitly define a measurement program. This model has been evaluated in terms of its applicability, usefulness, validity and reliability through a series of case studies. Currently, MEEGA seems to be used more widely in practice being reported by several studies from different authors evaluating different games and contexts [Calderón, 15].

Analyzing the quality factors used to evaluate educational games (RQ2), we observed that, there exist a large diversity of factors. However, the learning/knowledge improvement is in fact the factor most evaluated as expected as the main objective of educational games is to potentiate the students' learning. Learning is often evaluated by comparing the competence level after game playing with the competence level beforehand, typically based on a pre/post-test score [Mayer, 12] or through a self-assessment after game play [Fu, 09; Savi, 11]. Besides learning, most approaches also consider several other quality factors, such as challenge, competence, social interaction, fun, usability, etc. also confirming the findings of [Calderón, 15] with respect to educational games in diverse knowledge areas. These factors are evaluated as they are considered important in order to promote a deeper and active learning.

In general, we observed the lack of a consistent pattern of the quality factors to be evaluated and/or their decomposition. Some studies decompose the quality factors based literature, as [Savi, 11], explicitly decomposing motivation based on the ARCS model [Keller, 87], which defines four factors to represent motivation in instructional design: attention, relevance, confidence and satisfaction. On the other hand, some studies use the concept of motivation in a general way without any further decomposition [Carvalho, 12]. This can also be observed with respect to the concept of user experience. MEEGA [Savi, 11] decomposes user experience in terms of fun, competence, challenge, social interaction, and immersion, based on [Sweetser, 05; Poels, 07; Gámez, 09; Takatalo, 10]. On the other hand, others studies not explicitly decompose the concept of user experience [Carvalho, 12].

In general, we also observed a lack of methodological support provided in order to operationalize the data collection and analysis (RQ3). Only two approaches provide an explicit definition of the research design, data collection instruments, and data analysis methods.

Only one approach [Mayer, 12] proposes the usage of a more rigorous quasi-experimental research design as a best practice to assess game-based learning [All, 16]. One reason for the lack of further experimental research designs may be the effort required to conduct experiments by not only collecting data after the treatment but also before its application. This may cause a major disruption in the flow of the course and not well accepted by the learners themselves. In addition, experiments require control groups that may be impaired by using alternative instructional strategies being considered inferior. Furthermore, in order to obtain statistically significant results from such experiments a considerable sample size is required, especially when taking into consideration the need of not only the experimental but also a control group [All, 16]. Thus, even when undertaking this considerable amount of effort, the experimental studies may not yield significant results. A more viable alternative in practice may be the conduction of case studies, a non-experimental method, typically using a one-shot post-test only design, as proposed by MEEGA [Savi, 11]. Adopting this research strategy, the evaluation goal is assessed based on the students' perceptions through a standardized questionnaire after the game's application. An advantage is that the evaluation can be performed without a lot of effort in a relative non-intrusive way during the normal flow of the course. However, such self-assessments may lead to results with low validity, if data is collected via ad-hoc questionnaires or interviews. Therefore, a compromise may be the development of standardized questionnaires increasing the validity and reliability of the data being collected.

However, as result of our review we identified as a significant weakness the way how data collection instruments (typically questionnaires) are developed in an ad-hoc manner (RQ4). Yet, in order to obtain valid results, it is imperative to systematically define and operationalize the measures and data collection instruments [Kitchenham, 95; Kimberlin, 08]. Only two approaches [Savi, 11; Fu, 09] propose systematically developed and evaluated questionnaires. MEEGA has been developed by using GQM [Basili, 94], to explicitly define a measurement program [Savi, 11], systematically deriving analysis questions, measures and to guide the analysis of the collected. Both, MEEGA and EGameFlow use scale-development theory and methods proposed by [DeVellis, 03]. MEEGA and EGameFlow are also the only studies [Savi, 11; Fu, 09] that explicitly report a systematic evaluation (RQ5). The criteria used for validation [Savi, 11; Fu, 09], are defined based on scale-development theory [DeVellis, 03] including applicability, utility, validity, and reliability. The other approaches we encountered in our review seem to have been evaluated through case or pilot studies only not validating the models/data collection instruments [Carvalho, 12; Mayer, 12; Connolly, 09; Freitas, 06].

With respect to data analysis methods, most approaches also do not provide support. Again, only MEEGA and EGameFlow [Savi, 11; Fu, 09] provide explicit support on how to analyze the collected data. These approaches, typically, use quantitative methods, including descriptive statistics to measure central tendency, dispersion, and measures of dependency. To assist in the understanding results, graphical visualization techniques are used. In addition, EGameFlow also proposes the usage of hypothesis testing to compare two sample means [Fu, 09].

8.1 Threats to validity

As in any systematic review, there exist threats to validity to the results presented. We, therefore, identified potential threats and applied mitigation strategies in order to minimize their impact on our research.

Publication bias. Systematic reviews suffer from the common bias that positive outcomes are more likely to be published than negative ones [Kitchenham, 10]. Nevertheless, we do not consider this an essential threat to our research as rather than focusing on articles that present findings on the impact of these games, we aim at eliciting on how these games have been systematically evaluated.

Identification of studies. Another risk is the omission of relevant studies. In order to mitigate this risk, we carefully constructed the search string (see Table 5) in order to be as inclusive as possible considering not only core concepts but also synonyms. The risk of excluding relevant studies is further mitigated by the use of multiple databases, which cover the majority of scientific publications in the field.

Study selection and data extraction. Threats to study selection and data extraction have been mitigated with a detailed definition of the inclusion/exclusion criteria. We defined and documented a rigid protocol for the study selection and both authors conducted the selection together always discussing the selection until consensus was achieved.

9 Conclusions and Future Work

In this article, we present the state of the art on how to systematically evaluate educational games. We identified 11 articles, describing 7 different approaches to evaluate educational games. Most of them are frameworks rather than comprehensive evaluation methods, indicating a lack of support on how to conduct such evaluations. The encountered approaches also vary largely in terms of the quality factors evaluated. Besides evaluating the learning effect, they also consider challenge, usability, social interaction, etc. showing that there does not exist a pattern of the factors to be evaluated. Most of the approaches also seem to be developed in a rather ad-hoc manner, not providing an explicit definition of the objective, measures or data collection instruments.

In this respect, two approaches stand out: MEEGA and EGameFlow. Both approaches have been systematically developed by explicitly decomposing evaluation goals into measures and defining a questionnaire, evaluated through series of case studies. Both approaches focus on the evaluation of learning/knowledge improvement and user experience during the game play, including also in case of MEEGA the motivation promoted through the game. Currently, MEEGA seems to be used more widely in practice being reported by several studies from different authors evaluating different games and contexts, confirming also the findings of [Calderón, 15]. On the other hand, EGameFlow seems to have been applied so far only by the authors of the model themselves.

Based on the results of our review it becomes obvious that there exists a need for the identification of more consistent and uniform patterns for systematically evaluate educational games in order to obtain valid results that can be used to as a basis for decision on the application of such games and/or their continuous improvement.

Acknowledgements

This work was supported by the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico – www.cnpq.br), an entity of the Brazilian government focused on scientific and technological development.

References

[Abt, 02] Abt, C.C.: Serious Games. Lanhan, MD: University Press of America, 2002.

[Admiraal, 11] Admiraal, W., Huizenga, J., Akkerman, S., Dam, G. T.: The concept of flow in collaborative game-based learning, Computers in Human Behavior, 27(3), 1185–1194, 2011.

[Ak, 12] Ak, O.: A Game Scale to Evaluate Educational Computer Games, Procedia - Social and Behavioral Sciences, 46, 2477-2481, 2012.

[Akili, 06] Akili, G. K.: A New Approach in Education. In D. Gibson, C. Aldrich, M. Prensky (Eds.), Games and Simulations in Online Learning: Research and Development Frameworks, pp. 1-20, Information Science Publishing, 2006.

[All, 16] All, A., Castellar, E. P. N., Looy, J. V.: Assessing the effectiveness of digital game-based learning: Best practices, Computers & Education, 92–93, 90-103, 2016.

[Alliger, 97] Alliger, G. M., Tannenbaum, S. I., Bennett, W., Traver, H., Shotland, A.: A Meta-Analysis of the Relations among Training Criteria, Personnel Psychology, 50(2), 341–358, 1997.

[Baba, 93] Baba, D. M.: Determinants of video game performance. Advances in Psychology, 102, 57–74, 1993.

[Backlund, 13] Backlund, P., Hendrix, M.: Educational games - Are they worth the effort? A literature survey of the effectiveness of serious games, Proc. of the 5th Int. Conf. on Games and Virtual Worlds for Serious Applications, Poole, GB, 2013.

[Basili, 94] Basili, V. R., Caldiera, G., Rombach, H. D.: Goal, Question Metric Paradigm. In J. J. Marciniak, Encyclopedia of Software Engineering, pp. 528-532, Wiley-Interscience, New York, NY, USA, 1994.

[Battistella, 15] Battistella, P., Wangenheim, C. G.: ENgAGED: Games development process for Computing Education. Technical Report 01/2015, Brazilian Institute for Digital Convergence, Department of Informatics and Statistics, Federal University of Santa Catarina, Brazil.

[Battistella, 16] Battistella, P., Wangenheim, C. G.: Games for Teaching Computing in Higher Education – A Systematic Review. IEEE Technology and Engineering Education (ITEE) Journal, 9(1), 8-30, 2016.

[Berlyne, 60] Berlyne, D. E.: Conflict, arousal, and curiosity. New York: McGraw-Hill, 1960.

[Bloom, 56] Bloom, B. S.: Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain. New York: Toronto: Longmans, Green, 1956.

[Blumberg, 00] Blumberg, F. C.: The effects of children's goals for learning on video game performance, Journal of Applied Developmental Psychology, 21(6), 641–653, 2000.

[Boyle, 12] Boyle, E. A., Connolly, T. M., Hainey, T., Boyle, J. M.: Engagement in digital entertainment games: A systematic review. Computers in Human Behavior, 28(3), 771–780, 2012.

[Boyle, 16] Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., Pereira, J.: An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. Computers & Education, (94), 178-192, 2016.

[Branch, 10] Branch, R. M.: Instructional Design: The ADDIE Approach, Springer New York Dordrecht Heidelberg London, 2010.

[Calderón, 15] Calderón A., Ruiz M.: A systematic literature review on serious games evaluation: An application to software project management, Computers & Education, 87, 396-422, 2015.

[Carvalho, 12] Carvalho, C. V.: Is game-based learning suitable for engineering education?, Proc. of the Global Engineering Education Conf., IEEE, pp.1-8, 2012.

[Caulfield, 11] Caulfield, C., Xia, J., Veal, D., Maj, S. P.: A systematic survey of games used for software engineering education, Modern Applied Science, 5(6), 28-43, 2011.

[Clark, 03] Clark, M., Oxman, A. D.: Cochrane Reviewers' Handbook 4.2.0. Oxford: The Cochrane Library, 2003.

[Connolly, 07] Connolly, T. M., Stansfield, M., Hainey, T.: An application of games-based learning within software engineering, British Journal of Educational Technology, 38, 416-428, 2007.

[Connolly, 08] Connolly, T. M., Stansfield, M. H., Hainey, T.: Development a General Framework for Evaluating Games-based learning, Proc. of the 2nd European Conf. on Games-based Learning. Barcelona, Spain, 2008.

[Connolly, 09] Connolly, T. M., Stansfield, M. H., Hainey, T.: Towards the development of a games-based learning evaluation framework. In T. M. Connolly, M. H. Stansfield, & E. Boyle (Eds.), Games-based learning advancement for multisensory human computer interfaces: Techniques and effective practices, Idea-Group Publishing: Hershey, 2009.

[Connolly, 12] Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., Boyle, J. M.: A systematic literature review of empirical evidence on computer games and serious games, Computers & Education, 59(2), 661-686, 2012.

[Csikszentmihalyi, 91] Csikszentmihalyi, M.: Flow: The psychology of optimal experience. New York: Harper Perennial, 1991.

[Dempsey, 96a] Dempsey, J., Rasmussen, K., Lucassen, B.: The instructional gaming literature: Implications and 99 sources. Technical Report 96-1. College of Education, University of South Alabama, AL, 1996.

[Dempsey, 96b] Dempsey, J. V., Lucassen, B. A., Haynes, L. L., Casey, M. S.: Instructional applications of computer games. Paper presented at the American Education Research Association, NY, 1996.

[DeVellis, 03] DeVellis, R. F.: Scale development: theory and applications. SAGE Publications, 2003.

[Falchikov, 89] Falchikov, N., Boud, D.: Student Self-Assessment in Higher Education: A Meta-Analysis, Review of Educational Research, 59(4), 395–430, 1989.

[Fenton, 98] Fenton, N.E., Pfleeger, S.L.: Software Metrics: A Rigorous and Practical Approach, 2nd ed., PWS Pub. Co., Boston, MA, USA, 1998.

[Freedman, 07] Freedman, D., Pisani, R., Purves, R.: Statistics, 4th ed, New York: W. W. Norton & Company, 2007.

[Freitas, 06] Freitas, S. D., Oliver, M.: How can exploratory learning with games and simulations within the curriculum be most effectively evaluated?, Computers & Education, 46(3), 249-264, 2006.

[Fu, 09] Fu, F., Su, R.,Yu, S.: EGameFlow: A scale to measure learners' enjoyment of e-learning games, Computers & Education, 52(1), 101-112, 2009.

[Gámez, 09] Gámez, E. H. C.: On the Core Elements of the Experience of Playing Video Games. PhD Thesis. Interaction Centre Department of Computer Science, University College London, GB, 2009.

[Garris, 02] Garris, R., Ahlers, R., Driskell, J. E.: Games, Motivation, and Learning: A Research and Practice Model, Simulation Gaming, 33(4), 441-467, 2002.

[Gibson, 13] Gibson, B., Bell, T.: Evaluation of games for teaching computer science, Proc. of the 8th Workshop in Primary and Secondary Computing Education, pp. 51-60, ACM, New York, NY, USA, 2013.

[Gunter, 08] Gunter, G. A., Kenny, R. F., Vick, E. H.: Taking educational games seriously: using the RETAIN model to design endogenous fantasy into standalone educational games, Educational Technology Research and Development, 56, 511-537, 2007.

[Hainey, 10] Hainey, T., Connolly, T. M., Boyle, E. A.: A Refined Evaluation Framework for Games-based Learning, Proc. of the 4th European Conf. on Games-based Learning. Copenhagen, Denmark, 2010.

[Hamari, 16] Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., Edwards, T.: Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning, Computers in Human Behavior, 54, 170-179, 2016.

[Hays, 05] Hays, R.T.: The Effectiveness of Instructional Games: A Literature Review and Discussion, Naval Air Warfare Center Training System Division, Orlando, FL, USA, 2005.

[Herz, 97] Herz, J. C.: Joystick nation : how videogames ate our quarters, won our hearts, and rewired our minds. Little, Brown and Company, Boston, MA, 1997.

[Hevner, 10] Hevner, A., Chatterjee, S.: Design Research in Information Systems: Theory and Practice. Integrated Series in Information Systems, Springer, 2010.

[Keller, 87] Keller, J.: Development and Use of the ARCS Model of motivational Design, Journal of Instructional Development, 10(3), 2-10, 1987.

[Kiili, 05] Kiili, K.: Digital game-based learning: Towards an experiential gaming model, The Internet and Higher Education, 8(1), 13-2, 2005.

[Kimberlin, 08] Kimberlin, C. L., Winterstein, A. G.: Validity and reliability of measurement instruments used in research. Am J Health Syst Pharm, 65(23), 2276-84, 2008.

[Kirkpatrick, 06] Kirkpatrick, D. L., Kirkpatrick, J. D.: Evaluating training programs: the four levels, 3. ed., Berrett-Koehler Publishers, USA, 2006.

[Kitchenham, 95] Kitchenham, B., Pfleeger, S. L., Fenton, N.: Towards a Framework for Software Measurement Validation, IEEE Transactions on Software Engineering, 21(12), 929-944, 1995.

[Kitchenham, 10] Kitchenham, B.: Systematic literature reviews in software engineering – A tertiary study, Information and Software Technology, 52(1), 792-805, 2010.

[Martin, 08] Martin, A. J., Jackson, S.: Brief approaches to assessing task absorption and enhanced subjective experience: Examining short and core flow in diverse performance domains. Motivation and Emotion, 32(3), 141–157, 2008.

[Mayer, 12] Mayer, I.: Towards a Comprehensive Methodology for the Research and Evaluation of Serious Games, Procedia Computer Science, 15, 233-247, 2012.

[Mayer, 13] Mayer, I., Bekebrede, G., Warmelink, H., Zhou, Q.: A Brief Methodology for Researching and Evaluating Serious Games and Game-Based Learning. In Psychology, Pedagogy, and Assessment in Serious Games, Publisher: IGI Global, Editors: Connolly, Thomas M and Boyle, Liz and Hainey, Thomas and Baxter, Gavin and Moreno-Ger, Pablo, pp.357-393, 2013.

[Mayer, 14] Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., Ruijven, T., Lo, J., Kortmann, R., Wenzler, I.: The research and evaluation of serious games: Toward a comprehensive methodology, British Hournal of Educational Technology, 45(3), 502-527, 2014.

[Mayes, 01] Mayes, D. K., Cotton, J. E.: Measuring engagement in video games: A questionnaire. Proc. of the Human Factors and Ergonomics Society 45th Annual Meeting, pp. 692–696, 2001.

[Mitamura, 12] Mitamura, T., Suzuki, Y., Oohori T.: Serious Games for Learning Programming Languages, IEEE International Conference on Systems, Man, and Cybernetics, COEX, Seoul, Korea, 2012.

[Olsen, 11] Olsen, T., Procci, K., Bowers, C.: Serious games usability testing: How to ensure proper usability, playability, and effectiveness. In A. Marcus (Ed.), Design User Experience and Usability Theory Methods Tools and Practice Proc. First International Conf., 2011.

[Omar, 08] Omar, H.M., Jaafar, A.: Playability Heuristics Evaluation (PHE) approach for Malaysian educational games, Proc. of Int. Symposium on Information Technology, 3, pp.1-7, 2008.

[Omar, 10] Omar, H., Jaafar, A.: Heuristics evaluation in computer games, Proc. of the Int. Conf. on Information Retrieval & Knowledge Management, pp.188-193, 2010.

[Oslin, 98] Oslin, J. L., Mitchell, S. A., Griffin, L. L.: The game performance assessment instrument (GPAI), some concerns and solutions for further development, Journal of Teaching in Physical Education, 17(2), 220–240, 1998.

[Pfahl, 01] Pfahl, D., Ruhe, G., Koval, N.: An Experiment for Evaluating the Effectiveness of Using a System Dynamics Simulation Model in Software Project Management Education, Proc. of the 7th Int. Symposium on Software Metrics, pp.97-109, London, GB, 2001.

[Pinelle, 08] Pinelle, D., Wong, N., Stach, T.: Heuristic evaluation for games: usability principles for video game design. Proc. of the Conf. on Human Factors in Computing Systems, ACM, New York, NY, USA, 1453-1462, 2008.

[Poels, 07] Poels, K., Kort, Y. D., Ijsselsteijn, W.: It is always a lot of fun!: exploring imensions of digital game experience using focus group methodology. Proc. of Conf. on Future Play, pp. 83-89, Toronto, Canada, 2007.

[Preece, 02] Preece, J., Rogers, Y., Sharp, E.: Interaction Design: Beyond Human-computer Interaction, New York, NY: John Wiley & Sons, 2002.

[Prensky, 07] Prensky, M.: Digital Game-Based Learning. New York: Paragon House, 2007.

[Reichlin, 11] Reichlin, L., Mani, N., McArthur, K., Harris, A. M., Rajan, N., Dacso, C. C.: Assessing the acceptability and usability of an interactiveserious game in aiding treatment decisions for patients with localized prostate cancer. Journal of Medical Internet Research, 13(1), e4, 2011.

[Rodriguez-Cerezo, 14] Rodriguez-Cerezo, D., Sarasa-Cabezuelo, A., Gomez-Albarran, M., Sierra, J-L.: Serious games in tertiary education: A case study concerning the comprehension of basic concepts in computer language implementation courses. Computers in Human Behavior, 31, 558-570, 2014.

[Ross, 98] Ross, J. A., Rolheiser, C., Hogaboam-Gray, A.: Skills Training Versus Action Research InService: Impact on Student Attitudes to Self-Evaluation, Teaching and Teacher Education, 14(5), 463–77, 1998.

[Ross, 06] Ross, J. A.: The reliability, validity, and utility of self-assessment, Practical Assessment, Research & Evaluation, 11(10), 1-13, 2006.

[Savi, 11] Savi, R., Wangenheim, C. G., Borgatto, A. F.: A Model for the Evaluation of Educational Games for Teaching Software Engineering. Proc. of the 25th Brazilian Symposium on Software Engineering, pp. 194-203, São Paulo, Brazil, 2011 (in Portuguese).

[Schuurink, 08] Schuurink, E., Houtkamp, J., Toet, A.: Engagement and EMG in serious gaming: experimenting with sound and dynamics in the levee patroller training game. In P. Markopoulos, B. de Ruyter, W. Ijsselsteijn, & D. Rowland (Eds.), Fun and games: Second international conference, pp. 139-149, Eindhoven, The Netherlands: Springer Verlag, 2008.

[Seymour, 00] Seymour, E., Wiese, D., Hunter, A., Daffinrud, S. M.: Creating a better mousetrap: On-line student assessment of their learning gains. Paper presented at the National Meeting of the American Chemical Society, San Francisco, CA, 2000.

[Shadish, 02] Shadish, W. R., Cook, T. D., Campbell, D. T.: Experimental and quasi-experimental designs for generalized causal inference. New York: Houghton Mifflin Company, 2002.

[Sherry, 06] Sherry, J. L., Lucas, K., Greenber, B. S., Lachlan, K.: Video game uses and gratifications as predictors, Playing video games: motives, responses, and consequences, pp. 213-224, Mahwah, N.J.: Lawrence Erlbaum Associates, 2006.

[Sindre, 03] Sindre, G., Moody, D.: Evaluating the Effectiveness of Learning Interventions: an Information Systems Case Study, Proc. of the 11th European Conf. on Information Systems, Paper 80, Naples, Italy, 2003.

[Sitzmann, 10] Sitzmann, T., Ely, K. Brown, K. G., Bauer, K. N.: Self-Assessment of Knowledge: A Cognitive Learning or Affective Measure?, Academy of Management Learning & Education, 9(2), 169-191, 2010.

[Squires, 99] Squires, D., Preece, J.: Predicting quality in educational software: Evaluating for learning, usability and the synergy between them, Interacting with Computers, 11(5), 467-483, 1999.

[Soflano, 15] Soflano, M., Connolly, T. M., Hainey, T.: An application of adaptive games-based learning based on learning style to teach SQL. Computers & Education, 86, 192-211, 2015.

[Sweetser, 05] Sweetser, P., Wyeth, P.: GameFlow: a model for evaluating player enjoyment in games, Computers in Entertainment, 3(3), 1-24, 2005.

[Sweetser, 12] Sweetser, P., Johnson, D., Wyeth, P., Ozdowska, A.: GameFlow heuristics for designing and evaluating real-time strategy games. Proc. of the 8th Australasian Conf. on Interactive Entertainment: Playing the System, ACM, New York, NY, USA, Article 1, 10 pages, 2012.

[Takatalo, 10] Takatalo, J., Häkkinen, J., Kaistinen, J., Nyman, G.: Presence, Involvement, and Flow in Digital Games. In Bernhaupt, R. (Ed.). Evaluating User Experience in Games: Concepts and Methods, pp. 23-46, Springer, 2010.

[Tallir, 07] Tallir, I. B., Lenoir, M., Valcke, M., Musch, E.: Do alternative instructional approaches result in different game performance learning outcomes? Authentic assessment in varying game conditions, International Journal of Sport Psychology, 38(3), 263–282, 2007.

[Tang, 09] Tang, S., Hanneghan, M., El Rhalibi, A.: Introduction to Games-Based Learning. In Games-based Learning Advancement for Multisensory Human Computer Interfaces: Techniques and Effective Practices (Eds: T.M. Connolly, M.H. Stasnfield and E. Boyle), Idea-Group Publishing: Hershey, 2009.

[Topping, 03] Topping, K.: Self and Peer Assessment in School and University: Reliability, Validity and Utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), Optimising New Modes of Assessment: In Search of Qualities and Standards, 1, pp. 55–87, Dordrecht: Kluwer Academic Publishers, 2003.

[Trepte, 11] Trepte, S., Reinecke, L.: The pleasures of success: Game-related efficacy experiences as a mediator between player performance and game enjoyment, Cyberpsychology Behavior and Social Networking, 14(9), 555–557, 2011.

[Wangenheim, 09a] Wangenheim, C. G., Shull, F.: To Game or Not to Game? Software, IEEE, 26(2), 92-94, 2009.

[Wangenheim, 09b] Wangenheim, C.G., Kochanski, D., Savi, R.: Systematic Review on evaluation of games for software engineering learning in Brazil, Software Engineering Education Forum. Fortaleza, CE, Brazil, 2009 (in portuguese).

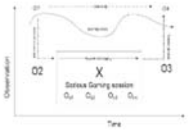
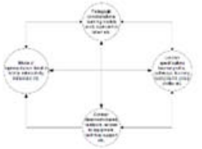
[Wangenheim, 12] Wangenheim, C. G., Wangenheim, A.: Ensinando Computação com Jogos. Florianópolis, SC, Brazil: Bookes, 2012.

[Wohlin, 12] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering, Springer-Verlag Berlin Heidelberg, 2012.

[Wouters, 13] Wouters, P., van Nimwegen, C., van Oostendorp, H., van der Spek, E. D.: A meta-analysis of the cognitive and motivational effects of serious games. Journal of Educational Psychology, 105(2), 249, 2013.

Appendix 1 – Data extracted to answer the questions RQ1 and RQ2

RQ1: Which models, methods, scales, or frameworks (approaches) exist to systematically evaluate educational games?					**RQ2:** Which quality and/or sub-quality factors are evaluated?		
Id	**Reference**	**Name**	**Instructional Strategy**	**Overview**	**Quality factors**	**Quality subfactors**	**Theoretical basis**
1	[Savi, 11]	MEEGA (Model for the Evaluation of Educational Games)	Educational games		Motivation User Experience Learning	Motivation: attention, relevance, confidence, satisfaction. User experience: fun, competence, challenge, social interaction, immersion. Learning	ARCS Model [Keller, 87] [Sweetser, 05; Poels07; Gámez, 09; Takatalo, 10] [Bloom, 56; Sindre, 03]
2	[Fu, 09]	EGameFlow	E-learning games	Not Informed	Concentration Goal clarity Feedback Challenge Control Immersion Social Interaction Knowledge improvement	Not Informed	[Sweetser, 05]
3	[Carvalho, 12]	Not defined	Game-based learning	Not Informed	Beta testing: - Game play - Game story - Mechanisms/Usability	Not Informed	Not Informed
		Not defined	Game-based learning	Not Informed	Gamma testing: - Knowledge - Motivation - Satisfaction	Knowledge Motivation: competence, interest, motivation for Computer Games Satisfaction: interest/enjoyment, perceived competence, user experience.	Not Informed
4	[Ak, 12]	Not defined	Educational computer games	Not Informed	Enjoyment Learning	Enjoyment: challenge, curiosity & mystery, clear goals, social interaction, diversion (fun), fantasy, arousal, flow. Learning	[Sherry, 06; Fu, 09; Garris, 02; Berlyne, 60; Kiili, 05; Csikszentmihalyi, 91; Freitas, 06; Squires, 99]

5	[Mayer, 12; Mayer, 13; Mayer, 14]	Not defined	Serious games	 O2 X O3 Serious Gaming session Observation Time	Game performance Game play Game experience Player satisfaction Learning	Game performance: time, avoidable mistakes Game play: dominance, influence, power Game experience: flow, immersion, presence Post-game: Game experience: engagement, fun Player satisfaction: clarity, relevance, attractiveness, ease of use, interaction with others students (social interaction), student's efforts, motivation, role identification, facilitator Learning: player learning satisfaction, self-reported, self-perceived learning, measured changes in knowledge, attitudes, skills, behaviors, asking clients, participants. Measured changes in team: safety, commitment, performance.	[Tallir, 07; Oslin, 98; Baba, 93; Trepte, 11; Blumberg, 00; Csikszentmihalyi, 91; Admiraal, 11; Martin, 08; Mayes, 11; Boyle, 12; Schuurink, 08; Olsen, 11; Reichlin, 11]
6	[Connolly, 09; Connolly, 08; Hainey, 10]	Evaluation Framework for Effective Games-based Learning	Game-based learning	Learner Performance Effective Games-based Learning GBL Environment Collaboration (Optional)	Learner performance/Learning Learner/academic motivation Learner/academic perceptions Learner/academic preferences GBL environment Collaboration between players where appropriate	Learner performance Learner/instructor motivation Learner/instructor perceptions Learner/instructor preferences Learner/instructor attitudes GBL environment: virtual environment, scaffolding, usability, level of social presence, deployment. Collaboration	Not clearly informed
7	[Freitas, 06]	Four Dimensional Framework	Games- and Simulation-based learning		Pedagogic considerations Learner specification Context Mode of representation	Pedagogic considerations: learning models used, approaches taken. Learner specification: learner profile, pathways, learning background, group profile. Context: classroom-based, outdoors, access to equipment, technical support Mode of representation: level of fidelity, interactivity, immersion.	Not Informed

Appendix 2 – Data extracted to answer the questions RQ3, RQ4 and RQ5

RQ3: How data collection and analysis is operationalized?					**RQ4:** How these approaches have been developed?	**RQ5:** How these approaches have been evaluated?				
Id	**Study type**	**Data collection instrument(s)**	**Response format**	**Data analysis method(s)**	**Development methodology**	**Evaluated factors**	**Number of applications**	**Data points**	**Data analysis method(s)**	**Findings**
1	Non-experimental with case study: one-shot post test only	Questionnaire	Likert scale Ordinal scale	Descriptive statistics: histogram, frequency diagram, median, mode.	GQM [Basili, 94] Scale Development [DeVellis, 03]	Applicability Utility Validity Reliability	Applied in two courses with 3 games in each application	79	Intercorrelation of scale items Item-total correlation Variance Mean Cronbach's alpha coefficient	The model was considered easy to use, requires little interruption of classes and the measuring instrument used for data collection presented satisfactory performance on a statistical analysis of validation.
2	Ad-hoc evaluation: pre-test/post-test	Questionnaire	Likert scale	Descriptive statistics: mean, SD, Pearson correlation coefficient Hypothesis testing: ANOVA, T-test Qualitative analysis	Scale Development [DeVellis, 03]	Item analysis Validity Reliability	One application in one course with 4 e-learning games	166	Mean Standard deviation Extreme group comparison Test for homogeneity T-Test ANOVA Pearson's correlation Cronbach's alpha coefficient	Statistical analyses showed that the scale developed in this study demonstrates high validity and reliability, which makes it an effective tool for assessing ‘‘the level of enjoyment brought to the learner by e-learning games.”
3	Ad-hoc evaluation: one-shot post-test only	Questionnaire Semi-structured interview	Likert scale	Not informed	Not informed	Not informed	Not informed	Not informed	Not informed	The initial results seem to indicate that the framework does in fact provides a formative view of the development process and allows establishing
	Ad-hoc evaluation: pre-test/post-	Questionnaire Tests	Likert scale Ordinal	Not informed	Not informed	Not informed	Applied in two classes with only one game	Not informed	Not informed	

	test		scale							the efficiency of the approach in what concerns the level of knowledge/skills achieved.
4	Not Informed	Not Informed	Not Informed	Not informed	Not informed	Not informed	Not Informed	Not Inform ed	Not informed	The expert analyzes and reliability and validity analyzes of the scale are not conducted yet, but the main structure of the scale is ready.
5	Quase-experimental: pre-test/post-test design	Not Informed	Not Informed	Not Informed	Not informed	Not informed	Several hundreds of sessions	2164	Not informed	We demonstrated the principles and workings of the model on the basis of a comparative case of twelve SGs.
6	Not Informed	Not Informed	Not Informed	Not Informed	Not informed	Not informed	Evaluated through 2 studies	Not inform ed	Not informed	Not informed
7	Not Informed	Not Informed	Not Informed	Not Informed	Not informed	Not informed	Applied to evaluate two games	Not inform ed	Not informed	This framework specifies the gap between the approaches and provides a tool which can help practitioners to bridge the two approaches, facilitating more critical and reflective process for embedding games and simulations in teaching practice.