Computers & Education 87 (2015) 396–422

Contents lists available at ScienceDirect

Computers & Education

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

A systematic literature review on serious games evaluation: An application to software project management

Alejandro Calderón*, Mercedes Ruiz

Department of Computer Science and Engineering, University of Cádiz, Avenida de la Universidad de Cádiz, 10, 11519, Puerto Real, Cádiz, Spain

ARTICLE INFO

Article history:
Received 7 May 2015
Received in revised form 20 July 2015
Accepted 21 July 2015
Available online 22 July 2015

Keywords:
Serious game
Evaluation
Systematic literature review
Software project management

ABSTRACT

Training that future practitioners receive in software project management is a topic of great importance. The objective of this systematic literature review is to summarize the current state of the art of the different methods and procedures used to assess serious games. The review follows a predefined procedure that involves automatically searching well-known digital databases. 1199 papers were found by the automatic searches in the digital databases and 102 papers were selected as primary studies. The process was complemented with manual searches using author and backward snowballing techniques. Our systematic literature review identified the main methods followed to assess serious games, the application domains in which the assessments took place, the categories of serious games assessed, the main features considered to assess the educational effectiveness of serious games, the procedures followed for the assessments and the size of the population that participated in the assessments. The results are useful to researchers and practitioners willing to assess serious games in different fields, but specially to those interested in assessing serious games in the area of software project management.

1. Introduction

Software project management is a knowledge area of great importance to keep in mind by experts on the road to success in the development of software projects. The relevance of teaching software project management for Information Technology (IT) learners has always been supported by organizations such as the Association for Computing Machinery (ACM) and IEEE-Computer Society in their joint task force curricula (ACM/IEEE-CS Joint Task Force on Computing Curricula, 2013, in which they recommend that learning about the methods and principles of project management should be practical rather than theoretical.

However, bringing professional practice into the classroom is a difficult task. Frequently, future computing professionals lack practical training with real-life scenarios. In addition, a highly theoretical learning environment, as in current software project management syllabi, is quite uninteresting for the future professionals (Ibrahim, 2011). Consequently, novel professionals develop their experience working in real projects, where the effects of a wrong plan or decision-making can lead to a failed project or the loss of benefit for the companies they work for.

* Corresponding author.
E-mail addresses: alejandro.calderon@uca.es (A. Calderón), mercedes.ruiz@uca.es (M. Ruiz).

http://dx.doi.org/10.1016/j.compedu.2015.07.011
0360-1315/

Given the importance that knowledge gained from practical and action-based learning experiences with real consequences have in software project management, the use of games and simulation-based experiences help us to achieve these objectives by teaching practically within a risk-free environment. Such games, designed with a different purpose than only entertainment, are called serious games (Abt, 2002; Zyda, 2005).

However, despite the potential benefits of using serious games in software project management, there is a lack of serious games to teach and assess learners in this area. The available games are frequently not flexible, have a very specific scope and do not allow assessing the learner's new skills automatically. The weaknesses observed in the available serious games led the authors of this work to develop a new simulation-based serious game that is able to bring real-life practice in software project management as well as performing an automatic assessment of the skills developed by the learners while playing the game (Calderón & Ruiz, 2014).

The next step, in our research work, is to perform the assessment of both the effectiveness of our game and its level of compliance with the objectives for which it was created. In order to design the assessment process of our game, we conducted this study aimed at finding the current state of the art of the different methods and procedures used to assess serious games, disregarding the application area of the serious game. In addition, the works found in the specific area of software project management were also identified and explicitly analyzed. The study was conducted as a systematic literature review (SLR) following the guidelines proposed by Kitchenham (Kitchenham & Charters, 2007; MacDonell, Shepperd, Kitchenham, & Mendes, 2010).

The contributions of this paper include: (i) summarizing the existing evidence on evaluation procedures and methods to assess serious games in different application domains, (ii) providing a quick reference for researchers interested in conducting serious games assessment, and (iii) providing grounded evidence of the assessment methods used to assess software project management serious games.

The structure of the paper is as follows. Section 2 presents the related work. In Section 3, the method used for the systematic literature review is introduced. Section 4 shows the results of the review, and Section 5 offers discussions on these results. In Section 6, describes the searches in the grey literature. Section 7 shows the main threats found to the validity of this study. Finally, Section 8 states conclusions and future work based on the findings obtained.

2. Related work

A systematic literature review is a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest. Individual studies contributing to a systematic review are called primary studies; a systematic review is a form of secondary study Kitchenham and Charters (2007).

This systematic literature review works in the field of serious games and their evaluation. Before attempting this work, we conducted an initial study to identify the existing secondary studies in the same scope. For that reason, we performed automatic searches in the following electronic databases: Scopus, IEEE Xplore, ACM Digital Library, SpringerLink and Web of Science. This search was conducted in October 2013, using the search string "(A1 **OR** A2 **OR** A3) **AND** B1 **AND** (C1 **OR** C2 **OR** C3 **OR** C4 **OR** C5 **OR** C6 **OR** C7)". Table 1 shows the search terms that are part of the search string.

When this search was performed, no significantly related secondary studies were identified. However, we found a systematic literature review in the scope of computer games and serious games and the empirical evidence of the potential positive impacts of gaming on users (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012). The study reviewed papers published between January 2004 and February 2009 to find evidence of the potential positive impacts of gaming on users aged 14 years or above, with respect to learning, skill enhancement and engagement. The paper concludes that even though empirical evidence concerning the effectiveness of games-based learning was found, there is a need for more Randomized Control Trials to provide more rigorous evidence of their effectiveness. Although the scope of this study is closely related to ours, there is a significance difference between both of them. The main aim of Connolly et al.'s work was to find empirical evidence of playing games outcomes. This aim is translated in the search terms used that link terms for games in conjunction with terms for possible outcomes, impacts or effects of playing games. In our case, we are not searching for the outcomes of playing games but for the different ways that have been used to evaluate serious games.

During the course of this study, several alerts were set in the electronic databases to check if a new paper matching this search string was published. Using the alerts allowed us to identify another secondary study, in this case a systematic

Table 1
Search terms to identify related secondary studies.

A1. Evaluation	B1. Serious Game	C1. Systematic review
A2. Validation		C2. Research review
A3. Assessment		C3. Systematic overview
		C4. Systematic literature review
		C5. Systematic mapping
		C6. Mapping study
		C7. Systematic mapping study

mapping study (SMS), about serious games quality that was published in August 2014. Vargas, García-Mundo, Genero, and Piattini (2014) conducted a study with the aim of determining the current state of serious games quality initiatives. In their work, the authors aim to discover the current state of serious games quality initiatives and to identify gaps that deserve future rigorous investigation. The mapping study was developed following the guidelines for SMS provided by Kitchenham and Charters (2007) and identifies the quality of serious games that have been addressed by researchers, mapping them with the quality features proposed in the ISO/IEC 25010 International Organization for Standardization (ISO) (2010). After analyzing this study, we decided to perform some manual searches in Google Scholar using the snowballing technique (Marshall, 1998) to observe if these authors had published some other works in this scope. These searches helped us identify a conference paper published in July 2014. García-Mundo, Vargas-Enríquez, Genero, and Piattini (2014) investigates whether the use of serious games in the field of computer science helps to improve learning, and analyzes what mechanisms and how they are used to assess the learning achieved.

Although our work shares similarities with the works above, the literature review presented in this paper is different because it focuses on serious games evaluation with the aim of identifying and classifying the evaluation methods used to evaluate serious games. For that reason, in our study we analyze other necessary and evaluation-related data that the works above did not analyze. In addition, our SLR covers a longer period of time analyzing publications until March 2015.

3. Method

This study has been undertaken as a systematic literature review based on the original guidelines proposed by Kitchenham and Charters (2007), which are among the most widely accepted in software engineering.

These guidelines establish that a review should comprise three phases: planning, conducting and reporting. In the planning activity, the review protocol is developed and how the researchers should work and interact to conduct the review is decided. This protocol defines the procedure for executing the review and includes research questions, search and evaluation strategies, inclusion/exclusion criteria, quality assessment, data collection form and methods of analysis. The second phase focuses on executing the protocol as it has been defined. Finally, the aim of the last phase is to elaborate the final report.

The main purpose of our work is to identify useful information about evaluation methods to assess serious games. Hence the goals of this systematic literature review are:

- Classify the application domains in which serious games assessments have taken place.
- Identify the categories of serious games in which the assessment experiences have been performed.
- Collect and analyze the existing procedures, methods and techniques used to assess serious games.

In the following sub-sections, we describe the activities carried out in each phase of this systematic literature review.

3.1. Research questions

In order to achieve the goals mentioned above, seven research questions have been set. These questions help to collect all the information needed to analyze the different evaluation. The research questions addressed by this study are:

- **RQ1**. What are the application domains in which serious games have been assessed?
- **RQ2**. What are the types of serious games that have been assessed within the former domains?
- **RQ3**. What methods, techniques or quality models have been used to assess these serious games?
- **RQ4**. What are the quality characteristics that have been assessed?
- **RQ5**. How are the evaluation models, techniques or methods applied to assess a serious game?
- **RQ6**. What is the size of the population involved in the existing assessment experiences of the serious games?

These questions can be classified into two areas of interest. On the one hand, questions RQ1 and RQ2 study the evaluated serious games and their features. These questions allow us to identify the application domain, the type and the knowledge area of the serious games evaluated by the evaluation methods.

On the other hand, RQ3, RQ4, RQ5, and RQ6 focus on the evaluation methods and their characteristics. To address RQ3, we analyzed the evaluation mechanisms that are used to assess serious games and we explore the models or frameworks that support these mechanisms. With respect RQ4, we observed the different characteristics that authors assess in their evaluations of serious games. With respect RQ5, we identified the performed procedures that authors follow in the serious games evaluation session. Finally, RQ6 shows the size of the population that takes part in the evaluation session of the serious games.

3.2. Search strategy

The aim of the search strategy is to identify the primary studies. An exhaustive search for papers was carried out to answer the proposed research questions.

First, we selected the keywords for the search. For this reason, general terms were used with the aim of assuring that most of the relevant research papers were included in the study. The main search terms were "Serious Game" and "Evaluation". The search string was constructed using the steps described in (Brereton, Kitchenham, Budgen, Turner, & Khalil, 2007), which are the following:

- Derive major terms from the questions by identifying the main concepts.
- Identify alternative spellings and synonyms for major terms.
- Check the keywords in any relevant papers we already have.
- Use the Boolean OR to add alternatives spellings and synonyms.
- Use the Boolean AND to link the major terms.

We performed some initial pilot searches to test and tune the search string, the final search strings consisted of the following Boolean expression "(A1 OR A2 OR A3) AND (B1 OR (B2 AND (C1 OR C2 OR C3)))" where search expressions are represented in Table 2.

The search was performed in the following digital databases: ACM Digital Library, IEEE Xplore, ISI Web of Science, SCOPUS and SpringerLink. We used MS Excel™ to store the information collected about the searches and Mendeley as the reference manager tool.

The researcher responsible of the review adapted the search string to each individual database. The searches were restricted to title, abstract, and keywords.

3.3. Study selection and the inclusion and exclusion criteria

The study selection process was performed by the first author of the paper, through a test-retest approach, and it was accomplished in the following two phases:

- **Phase 1.** Papers found during the search process were evaluated for their suitability based on the analysis of their title and abstract. The title and abstract of each article were reviewed against the inclusion and exclusion criteria (see Section 3.3). In this phase, publications that were clearly irrelevant were excluded. During this phase of the selection process, found papers were classified as possible select (PS) or non-selected paper (NS).
- **Phase 2.** Publications identified as possible select during Phase 1 were exposed to a more thorough analysis that included reading the conclusions or even the full text. This phase was done to ensure that the publication in question definitely contained information that is relevant to the study.

During this process the PhD supervisor, second author of this paper, selected a random sample of the selected papers to validate that the selection process was undertaken by an appropriate way.

This SLR identifies the papers that introduce an evaluation model, technique or method to assess serious games, explain the evaluation process of a serious game, or show any case of study about the application of an evaluation method to assess a serious game. The papers must be written in English and must have been published until March 2015. In order to get information about a large collection of serious games evaluation methods, we have not limited the start of the publication period. The systematic review protocol selected explicitly defines inclusion and exclusion criteria, as it is shown in Table 3.

Table 2
Search expressions.

A1. Evaluation	B1. Serious Game	C1. Education
A2. Validation	B2. Simulation Game	C2. Teaching
A3. Assessment		C3. Training

Table 3
Inclusion/Exclusion criteria.

Inclusion criteria	▪ Publications that present a model, technique or method to evaluate serious games. ▪ Articles that explain the assessment process of a serious game. ▪ Publications that show cases of study about the application of an evaluation model, technique or method to assess serious games.
Exclusion criteria	▪ Articles whose main aim is not evaluation methods for serious games. ▪ Articles that show a serious game, but do not provide information about its evaluation. ▪ Publications that present the outcomes of assessing a serious game but do not show any information about the evaluation technique used. ▪ Papers that discuss about evaluation models out of the serious games scope. ▪ Publications will be excluded, if only the abstract but not the full text is available. ▪ Articles not written in English will be excluded. ▪ Duplicate papers (same paper retrieved from different databases). ▪ Duplicate report of the same study (when several reports of a study exist in different databases, the most complete version of the study was included in the review).

3.4. Quality assessment

Each publication was evaluated for quality at the same time as the publication data extraction process was performed. A questionnaire, which had to be filled in for each included paper, was elaborated as quality instrument. The assessment questionnaire consists of nine quality assessment (QA) questions and is divided into two parts. The first part has questions to identify the quality of the paper regarding the main topic of this review (QA1-QA4); the questions of the second part aim to identify the quality of the information that the paper provides us, in order to know the relevance of the paper within the SLR (QA5-QA9). According with the above information, the nine quality assessment (QA) questions used are the following:

- **QA1**. Does the paper show an evaluation method?
- **QA2**. Does the paper show the information about the evaluation technique used?
- **QA3**. Does the paper provide the information about the models that support the evaluation method?
- **QA4**. Does the paper use examples to show the evaluation method application?
- **QA5**. Does the paper allow finding the appropriate information to answer RQ1?
- **QA6**. Does the paper allow answering RQ4?
- **QA7**. Does the paper allow finding the appropriate information to answer RQ5?
- **QA8**. Does the paper allow finding the appropriate information to answer RQ6?
- **QA9**. Does the paper allow finding the appropriate information to answer RQ2?

Each question was answered YES (Y) or NO (N). The questions were scored as follows:

- QA1: Y, the inclusion criteria are explicitly defined in the study, i.e., the paper presents an evaluation method to assess serious games, explains the evaluation process of a serious game, or shows any case of study about the application of an evaluation method to assess a serious game. N, the inclusion criteria are not explicitly defined and cannot be readily inferred.
- QA2: Y, the publication provides the information about the tools used in the evaluation method, and the questions of the questionnaire, if the evaluation method is based on the use of questionnaires. N, the publication does not show the required information.
- QA3: Y, the paper provides the models or the framework that support the evaluation technique. N, the publication does not present any information about those models.
- QA4: Y, the publication describes how the evaluation method is applied or provides an example of use. N, the paper does not provide any example of use.
- QA5: Y, the paper shows the application domain of the evaluated serious game (i.e. the paper answers RQ1). N, the application domain of the evaluated serious game is not defined and cannot be readily inferred.
- QA6: Y, the paper describes the assessed features or the quality characteristics of the evaluated serious games, or shows what features the method assesses (i.e. the paper answers RQ4). N, the features or the quality characteristics of the serious games that the method addresses are not defined and cannot be readily inferred.
- QA7: Y, the paper shows the procedure used to perform the evaluation session of a serious game (i.e. the paper answers RQ5). N, the paper does not show the procedure used to perform the evaluation session of a serious game; the presented procedure is incomplete and cannot be readily inferred.
- QA8: Y, the paper specifies the population that takes part into the evaluation process of an educational game (i.e. the paper answers RQ6). N, the information about the population does not appear in the paper and cannot be readily inferred.
- QA9: Y, the paper provides the information needed to know the type of serious game assessed. N, the type of serious game assessed is not defined and cannot be readily inferred.

The scoring procedure was Y = 1 and N = 0. Hence the total number of Ys defines the quality assessment score of each publication.

3.5. Data collection

From each paper, we extracted some general data regarding the electronic database where the study was retrieved and the date of retrieval, its title, authors, date and place of publication, abstract, language and whether the full text is available. Besides, more specific data needed to achieve our objectives were collected such as: a) the aims of the study, b) the answers to our research questions that the study provides, and c) the result of applying our quality assessment to the paper.

The first author of this paper collected the data for the publications included in the study. The PhD supervisor reviewed this process to ensure that the obtained data was appropriate. To assure that the data requirements were fully satisfied and the data obtained was accurate, papers were read completely in this phase. All collected data was stored within a spreadsheet, ensuring that all information related to the study was stored in one location. Furthermore, this also allowed us to easily compare and analyze the extracted data during the synthesis process.

3.6. Data analysis

The data of the selected papers was classified according to the following criteria:

- The evaluated serious game (addressing RQ1).
- The application domain of the evaluated serious game (addressing RQ1).
- The scope of knowledge of the evaluated serious game (addressing RQ1).
- The type of game (addressing RQ2).
- The evaluation mechanisms used to assess an educational game (addressing RQ3).
- Models or frameworks that support the evaluation mechanisms (addressing RQ3).
- To know is the paper provides, directly or indirectly, the questions of the evaluation methods, if the methods are based on any form of questionnaire (addressing RQ3).
- The evaluation features or quality characteristics of the educational games (addressing RQ4).
- The procedure to perform the evaluation session of the serious game (addressing RQ5).
- The size of the population that took part in the evaluation of the serious game (addressing RQ6).

4. Results

The SLR process was completed in sixteen months. During this period, we defined the SLR protocol, identified and selected the primary studies, performed the data extraction and synthesis process, and, finally, reported the results of our study. All the SLR process was conducted by two researchers: the first author of this paper performed the SLR process and the second reviewed the work of the first one. The SLR began in November 2013 with the development of the protocol for the SLR. The outline of the SLR process is shown in Table 4.

The following sub-sections present the results of the SLR. These results focus on the search results that are analyzed and on the quality evaluation results that are shown.

4.1. Search results

Papers related to serious game assessment published until September 2013, were retrieved in November 2013. In this search, we found 885 papers. To ensure that the papers published during the ongoing process were not missed, we defined alerts in the digital libraries. Therefore, from November 2013 to March 2015, all the newly published papers found by the alerts in the electronic databases were added to the process of this systematic review. As a result, before, the data synthesis process began, the number of publications found increased to 1199 papers. After analyzing these papers, we excluded 348 duplicate ones (Fig. 1).

4.2. Selection results

The selection process began with 851 papers. This process had two steps. In the first step, we analyzed the title and abstract of each paper following the inclusion and exclusion criteria. After this first review, the list of publications was reduced to 376 papers. Then we performed the second review of the selection process, where the full text of papers was analyzed. As a result of this second review, we excluded 257 additional papers from our list. Our primary studies were formed by 119 papers, which were analyzed to extract the data. Fig. 1 shows the evolution of the list of publications during the selection process.

Table 4
Outline of the SLR process.

Month	Activities	Outcomes
November 2013	SLR protocol development	Draft of protocol
	SLR draft protocol revision and improvement	SLR protocol version 1
	Search process	List of publications
	Removal of duplicates	List of publications without duplicate papers
December 2013	SLR protocol revision and improvement	The selection process was modified. SLR protocol version 2
	Publications Selection First Review (analysis of title and abstract)	List of possible selected publications
	Publications Selection Second Review (analysis of full text)	List of selected publications (primary studies)
April 2014	SLR protocol revision improvement	The data extraction process was modified. SLR protocol version 3
June 2014	Data extraction process	Tables of data
	Study quality assessment	Primary studies quality evaluation
November 2014	SLR protocol revision improvement	The data synthesis process was modified. SLR protocol version 4
	Data synthesis	Research questions answered
February 2015		First version of the report
March 2015	Report the results	Second version of the report
April 2015		Final report

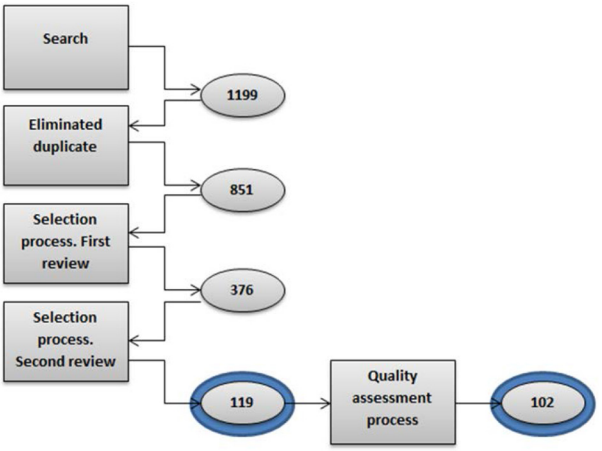

Fig. 1. Selection process.

Fig. 2 represents firstly, the papers that were retrieved from each consulted database; secondly, the number of different studies that were collected from each database after removing duplicates; thirdly, the number of studies that were collected from each database that passed the first review of the selection process. Finally, it shows the number of papers that were included as primary studies in this SLR.

In Fig. 3, we show the percentage of primary studies provided by each electronic database of the total number of primary studies accepted in our study. As expected, SCOPUS was the electronic that provided more primary studies to our SLR.

Fig. 4 shows the number of studies retrieved regarding the year of publication. It is noticeable the high number of publications published between 2009 and 2013, which is the year with the largest number of papers published. This allows us to see as the interest on serious games and its evaluation has been growing.

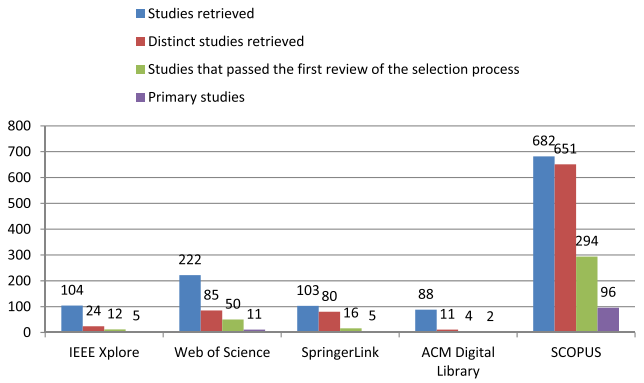

Fig. 2. Evolution of the studies retrieved in each electronic database.

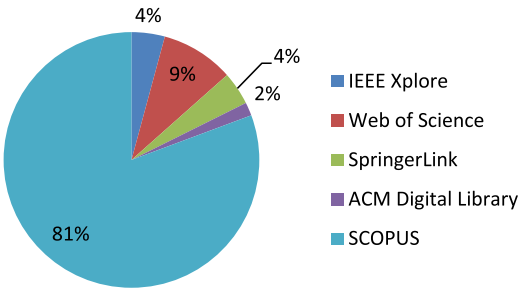

Fig. 3. Percentage of primary studies provided by each electronic database.

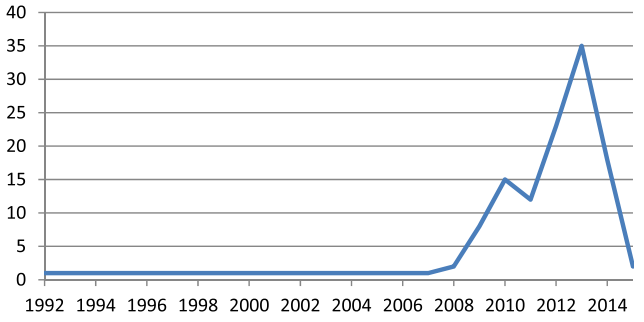

Fig. 4. Number of papers published regarding the year of publication.

Finally, Fig. 5 shows the number of selected papers regarding the year of publication and the electronic database. We can observe that the majority of selected papers are concentrated around the SCOPUS databases and above the year 2008.

4.3. Primary studies quality assessment

We assessed the finally selected papers (119) for quality using a questionnaire (see Section 3.4). This step allowed us to perform a third stage in the selection process based on detailed quality criteria. The nine quality questions were rated for each selected study in accordance with the criteria established in the quality questionnaire. Fig. 6 shows the coverage of every QA in the selected studies. It shows that QA1, QA4, QA5, QA6, QA7, QA8 and QA9 were covered in a rate higher than 70% by *Yes* answer. Moreover, more than a half of the QA questions were covered in a rate higher than 80% by *Yes* answer. In contrast, QA2

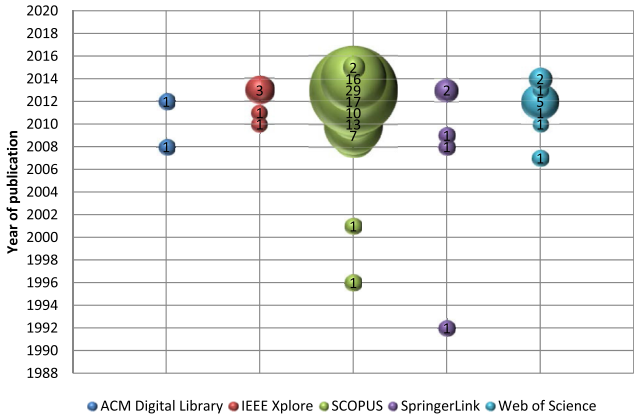

Fig. 5. Number of papers regarding the year of publication and the consulted databases.

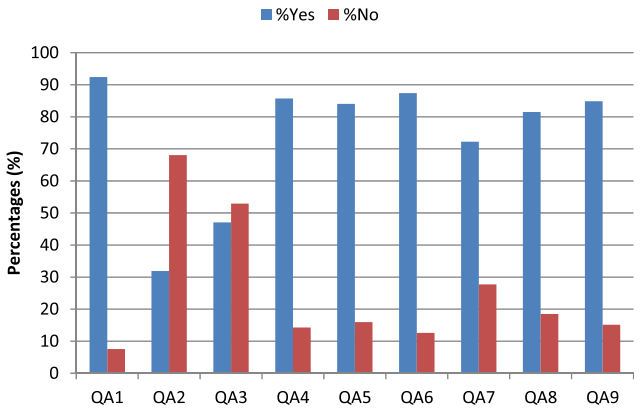

Fig. 6. Quality assessment results per question and type of assessment response.

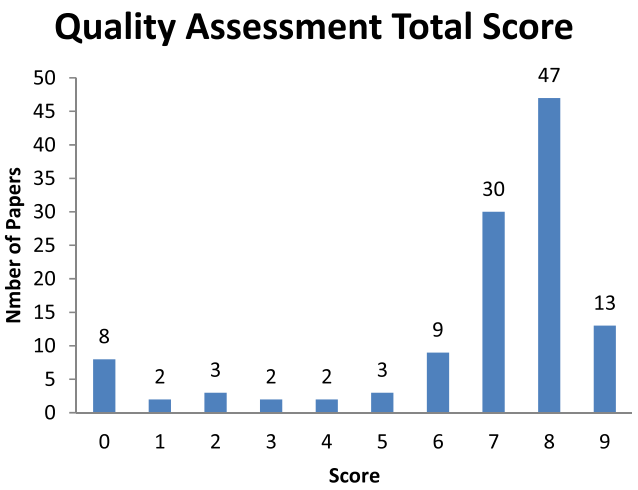

Fig. 7. Total score of papers' quality assessment.

and QA3 had a negative coverage. This is because most of the primary studies did not provide the required information about the set of questions of the evaluation method and nearly 50% of the primary studies did not provide enough information about the models used to create the evaluation method.

The score of the quality assessment have a value in a range from 0 (it does not provide enough information) to 9 (it provides all the required information). In Fig. 7, we summarize the total score obtained in the quality assessment and the number of primary studies that reached that score. Thereby, we can observe papers that reached a very low score in quality assessment, despite having passed the selection process. Specifically, there are 8 papers with a quality score of 0, which means that these papers do not provide any useful information related to our research questions. There are 9 papers with a quality score from 1 to 4, so these papers provide poor information relative to our research questions. The rest of the papers have a quality score greater than 5 and they were considered to provide the sufficient information to answer our research questions.

Fig. 8 compares the percentage of papers that have a score less than 5 with the percentage of papers with a score equal or greater than 5. The papers in the second group were the primary studies used in the synthesis process. We decided to take this decision in order to reduce the amount of null data, and also because the information provided by the papers with a 5-less score had also been found in the papers of the other group or it was no relevant. Finally, 102 papers took part in the synthesis process (see Table A1).

5. Discussion

In this section, we discuss our findings and the answers to the six research questions that were formulated.

RQ1. What are the application domains in which serious games have been assessed?

To answer this question, we analyzed the information about the application domains of the serious games in the primary studies. In these studies, more than ninety serious games, educational games and learning tools were evaluated within different domains. We have classified these application domains in six categories, as Table 5 shows:

Fig. 9 shows the total number of serious games in each application domain as well as the proportion of each domain according to the former classification. It can be observed that more than 50% of the games (53 games) are in the educational

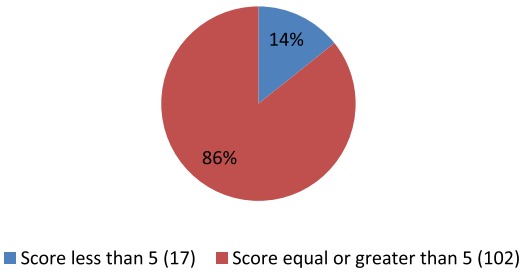

Fig. 8. Percentages of papers such as the score.

Table 5
Application domain categories.

Category	Definition	Primary studies
Health & wellness	This application domain includes serious games developed to improve people's quality life regarding their health and to create awareness about good habits during their daily life.	(Alamri et al., 2014; Backlund P. et al., 2011; Backlund P. et al., 2013; Berkovsky, Freyne, Coombe, Bhandari, & Baghaei, 2010; Bernardini, Porayska-Pomsta, & Smith, 2014; Burke et al., 2009; Buttussi & Chittaro, 2010; Cornejo, Hernandez, Tentori, & Favela, 2014; Enah, Piper, & Moneyham, 2014; Fuchslocher, Gerling, Masuch, & Kramer, 2011; Gerling, Fuchslocher, Schmidt, Kraemer, & Masuch, 2011; Hilborn, Cederholm, Eriksson, & Lindley, 2013; Khanal et al., 2014; Lányi, Brown, Standen, & Butkute, 2012; Lopez-Basterretxea, Mendez-Zorrilla, & Garcia-Zapirain, 2014; Lopez-Martinez et al., 2011; Manera et al., 2015; Petrasova, Cza, Chalmers, Farrer, & Wolke, 2010; Shah, Amirabdollahian, & Basteris, 2014; Yin, Ring, & Bickmore, 2012)
Culture	This application domain includes serious games for cultural training.	(Froschauer, Seidel, Gartner, Berger, & Merkl, 2010; Klinge, Rohmann, & Piontkowski, 2009; Manero, Fernandez-Vara, & Fernandez-Manjon, 2013; Mortara, Catalano, Fiucci, & Derntl, 2014; Rehm & Leichtenstern, 2012)
Professional learning & training	This application domain includes serious games that are used in companies to teach and train their staff.	(Buttussi et al., 2013; Cederholm, Hilborn, Lindley, Sennersten, & Eriksson, 2011; Chan, Qin, Chui, & Heng, 2012; Cowan et al., 2010; Diehl et al., 2013; Feldmann, Adam, & Bauer, 2014; Jarvis & Freitas, 2009; Kerga, Akaberi, Taisch, Rossi, & Terzi, 2013; Loreto, Mora, & Divitini, 2012; Pourabdollahian, Taisch, & Kerga, 2012; Qin, Chui, Pang, Choi, & Heng, 2010; Raybourn, 2009; Silva, Silva, Rossetti, & Leca Coelho, 2013; Tesei, Barbieri, Roceanu, & Beligan, 2014; Vidani, Chittaro, & Carchietti, 2010; Wiezer, Roozeboom, & Oprins, 2013; LeRoy Heinrichs, Youngblood, Harter, & Dev, 2008; Mayer, Wolff, & Wenzler, 2013)
Social	This application domain includes serious games for social skills training.	(Brown, Shopland, Baftersby, Lewis, & Evett, 2007; Coelho, Wesselius, & Papakonstantinou, 2010; Doukianou, Petridis, Dunwell, & Cooper, 2014; Rebolledo-Mendez, Avramides, de Freitas, & Memarzia, 2009)
Support	This application domain includes serious games created to support and help people in life's decisions.	(Dunwell et al., 2013)
Education	This application domain includes serious games created to teach, help, assess and motivate students in different areas of knowledge in formal education.	(Arnab, Perttula, & Suominen, 2014; Birchall & Gatzidis, 2011; Birchall & Gatzidis, 2013; El Borji & Khaldi, 2014; Brom, Šisler, & Slavík, 2010; Cheng, Su, Huang, & Chen, 2014; Costantino, Di Gravio, Shaban, & Tronci, 2012; Froschauer, Arends, Goldfarb, & Merkl, 2012; Hess, Wiemeyer, Hamacher, & Goesele, 2014; Ismailović, Pagano, & Brügge, 2011; Jaligama & Liarokapis, 2011; Kashibuchi & Sakamoto, 2001; Muratet, Torguet, Viallet, & Jessel, 2011; Paliokas, Arapidis, & Mpimpitsos, 2011; Quick, Corda, Chamberlin, Schaffner, & Byrd-Bredbenner, 2013; Wendel V., Gutjahr, Göbel, & Steinmetz, 2012; Wrzesien, López, & Raya, 2010; Zin & Yue, 2013; Mayer et al., 2013; Pohl, Rester, & Judmaier, 2009; Rodríguez-Cerezo, Sarasa-Cabezuelo, Gómez-Albarrán, & Sierra, 2014; Terzidou, Tsiatsos, Dae, Samaras, & Chasanidou, 2012; Adamo-Villani, Haley-Hermiz, & Cutler, 2013; Bascoul, Schmitt, Rasolofoarison, Chamberlain, & Lee, 2013; Basole, Bodner, & Rouse, 2013; Bevilacqua, Ciarapica, Emanuele, Mazzuto, & Paciarotti, 2014; Cook, McAloon, O'Neill, & Beggs, 2012; Cowley, Heikura, & Ravaja, 2013; Cowley, Fantato, Jennett, Ruskov, & Ravaja, 2014; Creutzfeldt, Hedman, Heinrichs, Youngblood, & Felländer-Tsai, 2013; Cronan, Léger, Robert, Babin, & Charland, 2012; Dudzinski, Greenhill, Kayyali, & Nabhani, 2013; Evans et al., 2015; Hannig, Kuth, Özman, Jonas, & Spreckelsen, 2012; Hauge, Duin, & Thoben, 2013; Hauge & Riedel, 2012; Hookham, Nesbitt, Cooper, Croft, & Rasiah, 2015; Hou & Li, 2014; Léger et al., 2012; Malliarakis, Satratzemi, & Xinogalos, 2014; Martin-Dorta et al., 2014; Parsons, Petrova, & Ryu, 2012; Qudrat-Ullah, 2010; Ranchhod, Gürău, Loukis, & Trivedi, 2012; Rughiniş, 2012; Couceiro, Papastergiou, Kordaki, & Veloso, 2013; Torrente, Moreno-Ger, Fernández-Manjón, & Del Blanco, 2009; Vaz de Carvalho, Lopes, & Gomes, 2013; Von Wangenheim, Thiry, & Kochanski, 2009; Von Wangenheim, Savi, & Borgatto, 2012; Wendel V. et al., 2013; Wildman & Reeves, 1996; Zhang, Caldwell, & Smith, 2013)

application domain. *Health & wellness* (20 games) and *Professional learning & training* (18 games) are the following categories, both representing about 20% of contributions of these domains.

In the case of educational games, we found that a larger number of the educational games that were evaluated are those that were used for training in higher education. Fig. 10 shows the proportion. The 60% of the educational serious games were evaluated in universities versus 40% of educational games that were evaluated in primary or secondary schools.

Secondly, since we have a special interest in educational serious games applied in the context of computer science, we deepened our analysis towards the games that have been evaluated in this scope. Only around 10% of the primary studies

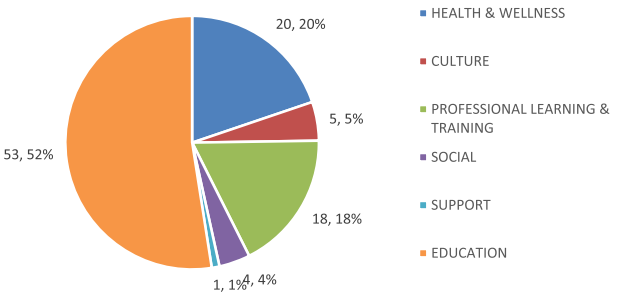

Fig. 9. Percentage of primary studies regarding the evaluated serious games per application domain.

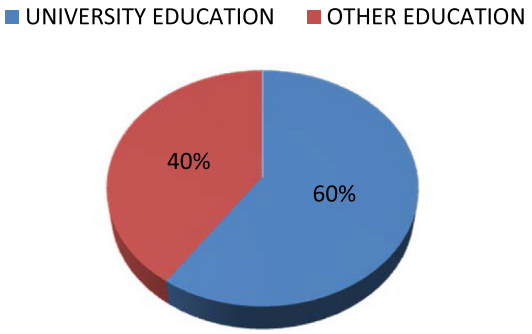

Fig. 10. Percentage of educational serious games that have been evaluated.

dealt with the evaluation of serious games in the computer science domain. Table 6 shows these primary studies classified by the different computer science topics they help training. Two primary studies assessed serious games related to the software project management area.

Programming is the computer science area that has attracted more interest in evaluating educational serious games. PlayLOGO (Paliokas et al., 2011) is a LOGO-like environment to implement Game Based Learning activities for children to learn the basic instances related to programming. It is an educational videogame that aims to familiarize students with the use of a programming language. In this game, students try to defeat one another in an interactive narrative applying LOGO commands as 'weapons'. Java Ninja (Zhang et al., 2013) is another game created to help students understand the concept of inheritance in Object-Oriented Programming (OOP). Players take the role of ninjas and they must use their programming knowledge to achieve new powers with the aim to solve the game challenges. Prog&Play (Muratet et al., 2011) aims also at strengthening players' programming skills. The game makes use of computer science metaphors, such as bits and pointers, calling them units. Units are represented by graphical objects which are controlled by the player Players give orders to their units to carry out operations through the Prog&Play Application Programming Interface (API), resulting in players giving their instructions through a program. In Adamo-Villani et al. (2013), the authors evaluate a serious game to teach 'Operator Precedence' to introductory programming students. Players move through a multilevel 3-dimensional maze and at each junction in the maze they are required to solve a mathematical problem that involves the application of operator precedence rules. A correct answer allows the player to move closer to the maze exit, an incorrect solution moves the player farther from the end of the maze. Robocode (El Borji & Khaldi, 2014) is another programming game, where the goal is to develop a robot battle tank to battle against other tanks in Java or.NET. CMX (Malliarakis et al., 2014) is an educational Massive Multiplayer Online Role Playing Game (MMORPG) that aims to introduce students to computer programming. This game helps players get acquainted with the way computer programs are structured and also allows them to engage in algorithmic logic. In Jaligama and Liarokapis (2011), an online virtual learning classroom environment focused for higher education is evaluated. In this virtual environment, students can login in the form of their avatars and follow lectures and laboratories in a collaborative

Table 6
Computer science topics.

Computer science topics	Primary studies
Programming	(Paliokas et al., 2011; Zhang et al., 2013; Muratet et al., 2011; Jaligama & Liarokapis, 2011; Adamo-Villani et al., 2013; Rodríguez-Cerezo et al., 2014; El Borji & Khaldi, 2014; Malliarakis et al., 2014)
Software project management	(Von Wangenheim et al., 2009; Von Wangenheim et al., 2012)
Operating systems	(Rughiniş, 2012)
Computer assembly	(Hou & Li, 2014)

manner. The authors test the feasibility of this serious game using the lecture material of the subject called '3D Graphics Programming'. Finally, 'Evaluators' is a system for the development of educational serious games oriented to introductory computer language implementation courses similar to those included in Computer Science tertiary curricula (Rodríguez-Cerezo et al., 2014). 'Evaluators' allows instructors to generate games from collections of exercises addressing basic concepts about the design and implementation of computer languages.

In the field of evaluation of serious games for software project management, we have only found two works. One of the works (Von Wangenheim et al., 2009) assesses a computer-based educational game prototype on software measurement called X-MED. The aim of the game is to exercise the application of software measurement in the context of project management. The game was developed for covering the cognitive levels remembering, understanding and applying in accordance to the revised version of Bloom's taxonomy of educational objectives. X-MED was designed as a single-player environment, in which player takes the role of a measurement analyst and defines and executes step-by-step a measurement program in a realistic scenario. The other work evaluates DELIVER!, a board game designed to help learners develop the skills needed to measure and control project performance by applying the Earned Value Management technique (Von Wangenheim et al., 2012). One of the main aims of this game is to motivate students in their learning process. The objective of the game is to be the first to deliver the results of the software project to the customer without running out of financial resources during execution. It is a game to be played in groups and covers the levels of remembering, understanding and applying of the revised version of Bloom's taxonomy of educational objectives.

There are also some other works addressed to evaluating serious games in other computer science related fields. In Rughiniş (2012) the authors evaluate the serious game 'World of USO' that is an online text game to teach Operating Systems based in Linux. The game is based on several types of activities that allow players get point. When a pre-established deadline is reached, the player with the highest score wins. Finally, 'Boom Room' is a problem-solving-based educational adventure game, developed for teaching knowledge of computer assembly that is evaluated in Hou and Li (2014). In the game, students must collect hardware and assemble a desktop computer to disable a bomb and escape a room within ten minutes.

RQ2. What are the types of serious games that have been assessed within the former domains?

During the analysis of the primary studies, we also collected the information about the type of serious game that the authors evaluated. We classified the serious games assessed on the primary studies in 8 categories as Table 7 shows.

Table 7
Game categories.

Type of game	Definition	Primary studies
Computer game	Serious game developed as computer software.	(Adamo-Villani et al., 2013; Alamri et al., 2014; Berkovsky et al., 2010; Bevilacqua, et al., 2014; Brom et al., 2010; Brown et al., 2007; Buttussi et al., 2013; Cederholm et al., 2011; Chan et al., 2012; Cheng et al., 2014; Coelho et al., 2010; Costantino et al., 2012; Couceiro et al., 2013; Cowan et al., 2010; Cowley et al., 2013; Cowley et al., 2014; Cronan et al., 2012; Doukianou et al., 2014; Dunwell et al., 2013; El Borji & Khaldi, 2014; Gerling et al., 2011; Hauge & Riedel, 2012; Hess et al., 2014; Hilborn et al., 2013; Hou & Li, 2014; Jarvis & Freitas, 2009; Khanal et al., 2014; Klinge et al., 2009; Lopez-Basterretxea et al., 2014; Lopez-Martinez et al., 2011; Lányi et al., 2012; Léger et al., 2012; Manera et al., 2015; Manero et al., 2013; Mayer et al., 2013; Mortara et al., 2014; Paliokas et al., 2011; Pavlas, Jentsch, Salas, Fiore, & Sims, 2012; Petrasova et al., 2010; Pohl et al., 2009; Qin et al., 2010; Qudrat-Ullah, 2010; Quick et al., 2013; Ranchhod et al., 2012; Raybourn, 2009; Rebolledo-Mendez et al., 2009; Rodríguez-Cerezo et al., 2014; Shah et al., 2014; Silva et al., 2013; Tesei et al., 2014; Vaz de Carvalho et al., 2013; Vidani et al., 2010; Von Wangenheim et al., 2009; Wendel V., Gutjahr, Göbel, & Steinmetz, 2012; Wendel V. et al., 2013; Wiezer et al., 2013; Yin et al., 2012; Zhang et al., 2013; Zin & Yue, 2013)
Mobile game	Serious game developed as a mobile application.	(Buttussi & Chittaro, 2010; Evans et al., 2015; Ismailović et al., 2011; Martin-Dorta et al., 2014; Parsons et al., 2012; Rehm & Leichtenstern, 2012)
Lego-based game	Serious game developed as computer software that uses Lego construction toys as part of the game.	(Kerga et al., 2013; Pourabdollahian et al., 2012)
Virtual worlds	Serious game developed as an interactive 3D virtual environment.	(Burke et al., 2009; Creutzfeldt et al., 2013; Froschauer et al., 2010; Hookham et al., 2015; Jaligama & Liarokapis, 2011; LeRoy Heinrichs et al., 2008; Terzidou et al., 2012; Wrzesien et al., 2010)
Web-based game	Serious game developed as a web application.	(Basole et al., 2013; Cook et al., 2012; Diehl et al., 2013; Dudzinski et al., 2013; Enah et al., 2014; Feldmann et al., 2014; Froschauer et al., 2012; Hannig et al., 2012; Hauge et al., 2013; Rughiniş, 2012)
Board game	Serious game developed as a board game.	(Bascoul et al., 2013; Kashibuchi & Sakamoto, 2001; Loreto et al., 2012; Von Wangenheim et al., 2012; Wildman & Reeves, 1996)
Videogame	Serious game that can be played in other platforms, such as videogame consoles (Xbox, PlayStation, etc.)	(Arnab et al., 2014; Backlund P. et al., 2011; Backlund P. et al., 2013; Bernardini et al., 2014; Birchall & Gatzidis, 2011; Birchall & Gatzidis, 2013; Cornejo et al., 2014; Muratet et al., 2011; Torrente et al., 2009)
MMORPG	Serious game developed as a Massively Multiplayer Online Role-Playing Game.	(Fuchslocher et al., 2011; Malliarakis et al., 2014)

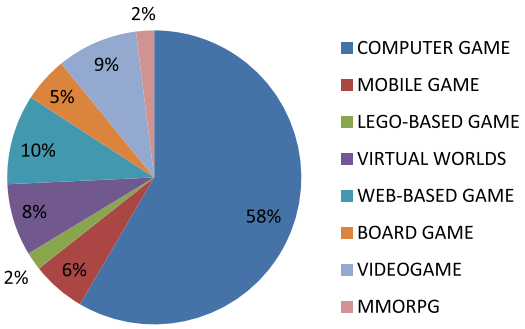

Fig. 11. Type of game.

Table 8
Evaluation techniques.

Technique	Definition	Primary studies
Questionnaires	The method uses forms of questions to assess the serious game.	(Adamo-Villani et al., 2013; Alamri et al., 2014; Arnab et al., 2014; Backlund P. et al., 2011; Backlund P. et al., 2013; Bascoul et al., 2013; Basole et al., 2013; Berkovsky et al., 2010; Bernardini et al., 2014; Bevilacqua et al., 2014; Birchall & Gatzidis, 2011; Birchall & Gatzidis, 2013; Brom et al., 2010; Brown et al., 2007; Burke et al., 2009; Buttussi & Chittaro, 2010; Buttussi et al., 2013; Cederholm et al., 2011; Cheng et al., 2014; Cook et al., 2012; Costantino et al., 2012; Couceiro et al., 2013; Cowan et al., 2010; Cowley et al., 2013; Cowley et al., 2014; Creutzfeldt et al., 2013; Cronan et al., 2012; Diehl et al., 2013; Doukianou et al., 2014; Dudzinski et al., 2013; Dunwell et al., 2013; Evans et al., 2015; Feldmann et al., 2014; Froschauer et al., 2012; Froschauer et al., 2010; Fuchslocher et al., 2011; Gerling et al., 2011; Hannig et al., 2012; Hauge et al., 2013; Hauge & Riedel, 2012; Hess et al., 2014; Hilborn et al., 2013; Hookham et al., 2015; Hou & Li, 2014; Kashibuchi & Sakamoto, 2001; Kerga et al., 2013; Khanal et al., 2014; Klinge et al., 2009; LeRoy Heinrichs et al., 2008; Lopez-Basterretxea et al., 2014; Lopez-Martinez et al., 2011; Loreto et al., 2012; Lányi et al., 2012; Léger et al., 2012; Malliarakis et al., 2014; Manera et al., 2015; Manero et al., 2013; Martin-Dorta et al., 2014; Mayer et al., 2013; Mortara et al., 2014; Muratet et al., 2011; Olsen, Procci, & Bowers, 2011; Paliokas et al., 2011; Parsons et al., 2012; Pavlas et al., 2012; Petrasova et al., 2010; Pourabdollahian et al., 2012; Qin et al., 2010; Qudrat-Ullah, 2010; Quick et al., 2013; Ranchhod et al., 2012; Raybourn, 2009; Rebolledo-Mendez et al., 2009; Rehm & Leichtenstern, 2012; Rodríguez-Cerezo et al., 2014; Rughiniş, 2012; Shah et al., 2014; Terzidou et al., 2012; Tesei et al., 2014; Torrente et al., 2009; Vaz de Carvalho et al., 2013; Vidani et al., 2010; Von Wangenheim et al., 2009; Von Wangenheim et al., 2012; Wendel V., Gutjahr, Göbel, & Steinmetz, 2012; Wendel V. et al., 2013; Wiezer et al., 2013; Wildman & Reeves, 1996; Wrzesien et al., 2010; Yin et al., 2012; Zhang et al., 2013; Zin & Yue, 2013)
Interviews	The method uses forms of questions to assess the serious game in oral way.	(Rughiniş, 2012; Zin & Yue, 2013; Quick et al., 2013; Basole et al., 2013; Backlund P. et al., 2013; Costantino et al., 2012; Loreto et al., 2012; Terzidou et al., 2012; Parsons et al., 2012; Yin et al., 2012; Ismailović et al., 2011; Fuchslocher et al., 2011; Cederholm et al., 2011; Gerling et al., 2011; Backlund P. et al., 2011; Brom et al., 2010; Raybourn, 2009; Rebolledo-Mendez et al., 2009; Silva et al., 2013; Cornejo et al., 2014; Vaz de Carvalho et al., 2013)
Logs	The method uses the logs of information generated during the evaluation session to assess the serious game.	(Buttussi & Chittaro, 2010; Hauge & Riedel, 2012; LeRoy Heinrichs et al., 2008; Terzidou et al., 2012; Adamo-Villani et al., 2013; Dudzinski et al., 2013; Malliarakis et al., 2014)
Discussions	The method uses oral discussions to assess the serious game.	(Basole et al., 2013; Costantino et al., 2012; Froschauer et al., 2012; Terzidou et al., 2012; Pohl et al., 2009; Malliarakis et al., 2014; Enah et al., 2014)
Videos	The method uses video recordings to assess the serious game.	(Bernardini et al., 2014; Arnab et al., 2014; Enah et al., 2014; Loreto et al., 2012; Terzidou et al., 2012; Jaligama & Liarokapis, 2011)
Frameworks	The method uses a group of techniques and steps defined by the authors to assess the serious game.	(Brom et al., 2010)
Observations	The method is based on the observation of the session progress to assess the serious game.	(Brown et al., 2007; Terzidou et al., 2012; Ismailović et al., 2011; Jaligama & Liarokapis, 2011; Wrzesien et al., 2010; Coelho et al., 2010; Wildman & Reeves, 1996; Arnab et al., 2014)
Others Methods	The primary study uses a method different of the previous techniques. In this case, the first study uses a correlational methodology to assess the serious game. The second one uses an evaluation grid.	(Lopez-Martinez et al., 2011; El Borji & Khaldi, 2014)

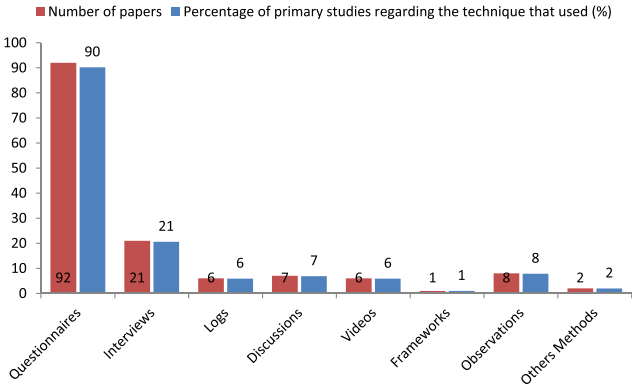

Fig. 12. Used techniques to evaluate a serious game regarding the papers that used its.

In Fig. 11, we present the percentages of primary studies regarding the category of the serious game that they evaluated. We can see as *Computer game* is the most frequent category with nearly 60% of the primary studies. Then *Web-based game* and *Videogame categories* are the second, both with nearly 10% of the primary studies. The last categories are *Lego-based game* with 2% of the primary studies and *MMORPG with 2% of the primary studies too*.

RQ3. What methods, techniques and quality models have been used to assess these serious games?

The aim of this question was to get an overview of the techniques or methods that authors used to evaluate serious games. Table 8 summarizes the techniques found.

In Fig. 12, we can observe that 90% of the primary studies used *questionnaires* as the main method for the evaluation of serious games. This technique is followed by *interviews*, which were used as the evaluation technique by approximately 20% of the primary studies. Questionnaires and interviews are by far the most used techniques to assess serious games. The other techniques are present only in less than 10% of the primary studies.

Given than the most used methods to assess serious games are based on a set of questions, we extended our analysis to the types of questions in these methods. Only 38% of the primary studies provided the questions used to collect the information to assess the serious game. Most of these studies used both quantitative and qualitative questions, being the Likert scale the preferred method to help users provide their answers to the questions. The primary studies that provide the questions used to assess the serious game are the following: (Alamri et al., 2014; Bascoul et al., 2013; Basole et al., 2013; Bevilacqua et al., 2014; Brom et al., 2010; Buttussi & Chittaro, 2010; Buttussi et al., 2013; Cook et al., 2012; Costantino et al., 2012; Couceiro et al., 2013; Cowley et al., 2013; Cowley et al., 2014; Creutzfeldt et al., 2013; Cronan et al., 2012; Dudzinski et al., 2013; Dunwell et al., 2013; El Borji & Khaldi, 2014; Hookham et al., 2015; Lányi et al., 2012; Léger et al., 2012; Lopez-Basterretxea et al., 2014; Malliarakis et al., 2014; Martin-Dorta et al., 2014; Mortara et al., 2014; Paliokas et al., 2011; Parsons et al., 2012; Pavlas et al., 2012; Qudrat-Ullah, 2010; Ranchhod et al., 2012; Rebolledo-Mendez et al., 2009; Rodríguez-Cerezo et al., 2014; Tesei et al., 2014; Torrente et al., 2009; Vaz de Carvalho et al., 2013; Vidani et al., 2010; Von Wangenheim et al., 2012; Wendel V., Gutjahr, Göbel, & Steinmetz, 2012; Zhang et al., 2013; Zin & Yue, 2013).

RQ4. What are the quality characteristics that have been assessed?

The aim of this research question is to observe the quality characteristics that the authors took into account to evaluate their serious games. We identified 18 different quality characteristics that were assessed with the aim of evaluating serious games. Table 9 describes these quality characteristics and show the primary studies that asses them.

Fig. 13 shows these quality characteristics and the number of primary studies that evaluated each of them. In this analysis, we can observe that nearly 50% of our primary studies assessed the learning outcomes of the serious game. This shows that the main interests when assessing the impact of serious games are to determine the games educational effectiveness in terms of knowledge acquisition. In addition, usability and user's experience are the following characteristics more frequently assessed. Regarding the priority that the authors have given to the assessment of theses quality characteristics, we can affirm that in most of the works, the assessment was aimed at verifying whether the games satisfied the objectives for which they had been created.

RQ5. How are the evaluation models, techniques or methods applied to assess a serious game?

This research question searches the steps that authors followed to perform the evaluation of the serious games. For this reason, we focused on the analysis of when the evaluation methods were used during the process of the serious game assessment. In this analysis, we found three main types of procedure which are defined in Table 10.

Table 9
Quality characteristics.

Quality characteristic	Definition	Primary studies
Game design	Design and aesthetic of the serious game.	(Basole et al., 2013; Cederholm et al., 2011; Lopez-Martinez et al., 2011; Muratet et al., 2011)
User's satisfaction	Attitude of the user towards the serious game.	(Alamri et al., 2014; Bascoul et al., 2013; Evans et al., 2015; LeRoy Heinrichs et al., 2008; Lopez-Basterretxea et al., 2014; Lopez-Martinez et al., 2011; Martin-Dorta et al., 2014; Mortara et al., 2014; Rughiniş, 2012; Torrente et al., 2009; Vidani et al., 2010)
Usability	Ease of use and learnability of the serious game.	(Adamo-Villani et al., 2013; Backlund P. et al., 2011; Backlund P. et al., 2013; Basole et al., 2013; Birchall & Gatzidis, 2011; Birchall & Gatzidis, 2013; Burke et al., 2009; Buttussi & Chittaro, 2010; Cederholm et al., 2011; Costantino et al., 2012; Couceiro et al., 2013; Cowan et al., 2010; Diehl et al., 2013; Doukianou et al., 2014; Froschauer et al., 2012; Froschauer et al., 2010; Gerling et al., 2011; Hannig et al., 2012; Hess et al., 2014; Hilborn et al., 2013; Hookham et al., 2015; Ismailović et al., 2011; Khanal et al., 2014; LeRoy Heinrichs et al., 2008; Loreto et al., 2012; Mortara et al., 2014; Muratet et al., 2011; Olsen et al., 2011; Paliokas et al., 2011; Petrasova et al., 2010; Raybourn, 2009; Rehm & Leichtenstern, 2012; Rodríguez-Cerezo et al., 2014; Vaz de Carvalho et al., 2013; Vidani et al., 2010; Zhang et al., 2013; Zin & Yue, 2013)
Usefulness	The utility of the serious game.	(Backlund P. et al., 2011; Backlund P. et al., 2013; Birchall & Gatzidis, 2011; Birchall & Gatzidis, 2013; Buttussi & Chittaro, 2010; Costantino et al., 2012; Vidani et al., 2010; Zhang et al., 2013)
Understandability	The ability of the serious game of being understood.	(Birchall & Gatzidis, 2013; Buttussi & Chittaro, 2010)
Motivation	The influence of the serious game in the user's motivation.	(Backlund P. et al., 2011; Birchall & Gatzidis, 2013; Buttussi & Chittaro, 2010; Kashibuchi & Sakamoto, 2001; Manera et al., 2015; Mortara et al., 2014; Muratet et al., 2011; Shah et al., 2014; Terzidou et al., 2012; Torrente et al., 2009; Von Wangenheim et al., 2012; Zhang et al., 2013)
Performance	The performance of the serious game.	(Cowley et al., 2013; Khanal et al., 2014; Manero et al., 2013; Qin et al., 2010; Zhang et al., 2013)
Playability	The ability of the serious game of being played.	(Basole et al., 2013; Burke et al., 2009; Diehl et al., 2013; El Borji & Khaldi, 2014; Manero et al., 2013; Olsen et al., 2011; Petrasova et al., 2010)
Pedagogigal aspects	Educational aspects of the serious game.	(El Borji & Khaldi, 2014; Hauge & Riedel, 2012; Manero et al., 2013)
Learning outcomes	What learners will know or be able to do as a result of playing the serious game.	(Berkovsky et al., 2010; Bernardini et al., 2014; Bevilacqua et al., 2014; Brown et al., 2007; Buttussi et al., 2013; Chan et al., 2012; Cheng et al., 2014; Cook et al., 2012; Couceiro et al., 2013; Cowan et al., 2010; Cowley et al., 2014; Cronan et al., 2012; Diehl et al., 2013; Dudzinski et al., 2013; Evans et al., 2015; Froschauer et al., 2012; Froschauer et al., 2010; Hannig et al., 2012; Hauge & Riedel, 2012; Kerga et al., 2013; Hookham et al., 2015; Hou & Li, 2014; Ismailović et al., 2011; Jaligama & Liarokapis, 2011; Jarvis & Freitas, 2009; Kashibuchi & Sakamoto, 2001; Loreto et al., 2012; Léger et al., 2012; Martin-Dorta et al., 2014; Mayer et al., 2013; Mortara et al., 2014; Muratet et al., 2011; Olsen et al., 2011; Paliokas et al., 2011; Parsons et al., 2012; Qudrat-Ullah, 2010; Quick et al., 2013; Ranchhod et al., 2012; Rughiniş, 2012; Rehm & Leichtenstern, 2012; Rodríguez-Cerezo et al., 2014; Terzidou et al., 2012; Tesei et al., 2014; Von Wangenheim et al., 2009; Von Wangenheim et al., 2012; Wiezer et al., 2013; Zin & Yue, 2013)
Engagement	The ability of the serious game to engage users.	(Adamo-Villani et al., 2013; Arnab et al., 2014; Brown et al., 2007; Cowan et al., 2010; Dunwell et al., 2013; Evans et al., 2015; Mortara et al., 2014; Petrasova et al., 2010; Pourabdollahian et al., 2012; Von Wangenheim et al., 2009; Wrzesien et al., 2010)
User's experience	User's behavior, attitude, and emotions using the serious game.	(Bascoul et al., 2013; Bevilacqua et al., 2014; Coelho et al., 2010; Cornejo et al., 2014; Cowley et al., 2014; Creutzfeldt et al., 2013; Dudzinski et al., 2013; Feldmann et al., 2014; Gerling et al., 2011; Hauge et al., 2013; Hess et al., 2014; Hookham et al., 2015; Hou & Li, 2014; Manera et al., 2015; Pavlas et al., 2012; Ranchhod et al., 2012; Raybourn, 2009; Silva et al., 2013; Von Wangenheim et al., 2012; Wendel V., Gutjahr, Göbel, & Steinmetz, 2012; Wendel V. et al., 2013; Wrzesien et al., 2010; Yin et al., 2012)
Efficacy	The ability of the serious game to produce the desired result.	(Adamo-Villani et al., 2013; Creutzfeldt et al., 2013; Khanal et al., 2014; Malliarakis et al., 2014; Mortara et al., 2014; Yin et al., 2012)
Social impact	The effects that the serious game produce in the population.	(Rebolledo-Mendez et al., 2009)
Cognitive behavior	The ability of the serious game to produce effects in user's cognitive behavior.	(Alamri et al., 2014; Hauge & Riedel, 2012; Manera et al., 2015; Quick et al., 2013)
Enjoyment	The ability of the serious game of providing a fun experience to the users.	(Backlund P. et al., 2013; Berkovsky et al., 2010; Fuchslocher et al., 2011; Jaligama & Liarokapis, 2011; Wrzesien et al., 2010)
Acceptance	How users receive the serious game.	(Backlund P. et al., 2011; Backlund P. et al., 2013; Brom et al., 2010; Enah et al., 2014; Feldmann et al., 2014; Fuchslocher et al., 2011; Hou & Li, 2014; Klinge et al., 2009; Loreto et al., 2012; Manera et al., 2015)
User interface	The interactions of the users with the serious game.	(Lányi et al., 2012)

The three types of procedure defined above are usually preceded by an introductory phase where the goals of the experiment and the serious game are introduced to the users. Demographics data about the users are also collected. Fig. 14 shows the number of primary studies that used each kind of procedure. We can observe how simple procedure was the most used.

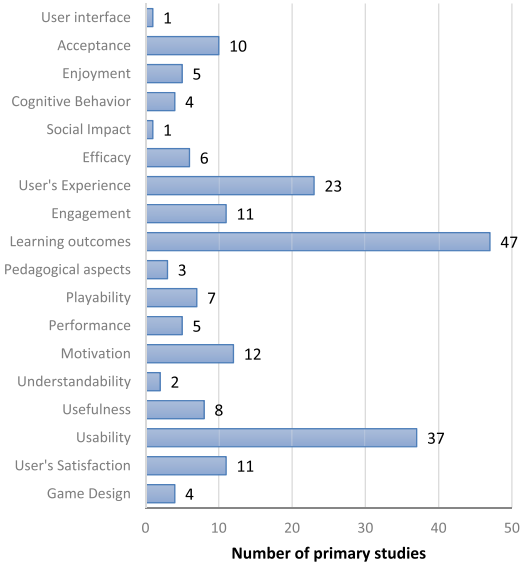

Fig. 13. Quality characteristics.

In this analysis, we observed that some authors use an experimental method based on control groups (Shuttleworth, 2010) to evaluate the effectiveness of the serious game by comparison. The use of two groups seeks to analyze the learning that users acquire using the serious game (experimental group) and using others methods (control group) to compare the two ways of learning and to observe if really the best results are achieved with the use of the serious game. Only 15 of our primary studies included the use of control and experimental groups in their procedures. These primary studies are the following:

Table 10
Types of procedure.

Procedure	Definition	Primary studies
Simple	In th; is procedure, authors carry out a session with the serious game and, after playing the game, the evaluation mechanisms are provided to the players.	(Adamo-Villani et al., 2013; Alamri et al., 2014; Arnab et al., 2014; Backlund P. et al., 2011; Basole et al., 2013; Burke et al., 2009; Buttussi & Chittaro, 2010; Cook et al., 2012; Costantino et al., 2012; Cowan et al., 2010; Creutzfeldt et al., 2013; Cronan et al., 2012; Doukianou et al., 2014; Dunwell et al., 2013; Feldmann et al., 2014; Fuchslocher et al., 2011; Gerling et al., 2011; Hauge et al., 2013; Hess et al., 2014; Hookham et al., 2015; Jaligama & Liarokapis, 2011; Jarvis & Freitas, 2009; Loreto et al., 2012; Léger et al., 2012; Malliarakis et al., 2014; Manera et al., 2015; Manero et al., 2013; Muratet et al., 2011; Olsen et al., 2011; Paliokas et al., 2011; Parsons et al., 2012; Pavlas et al., 2012; Petrasova et al., 2010; Pourabdollahian et al., 2012; Qin et al., 2010; Qudrat-Ullah, 2010; Ranchhod et al., 2012; Raybourn, 2009; Rebolledo-Mendez et al., 2009; Rehm & Leichtenstern, 2012; Rodríguez-Cerezo et al., 2014; Rughiniş, 2012; Shah et al., 2014; Terzidou et al., 2012; Torrente et al., 2009; Vaz de Carvalho et al., 2013; Vidani et al., 2010; Von Wangenheim et al., 2012; Wendel V., Gutjahr, Göbel, & Steinmetz, 2012; Wendel V. et al., 2013)
Pre/Post	This procedure is characterized by two stages of evaluation. One before using the serious game and another one after. This kind of procedure is usually used by authors who evaluated the level of knowledge acquisition that players acquired using the serious game.	(Brown et al., 2007; Backlund P. et al., 2013; Bascoul et al., 2013; Berkovsky et al., 2010; Bernardini et al., 2014; Bevilacqua et al., 2014; Birchall & Gatzidis, 2011; Birchall & Gatzidis, 2013; Brom et al., 2010; Cederholm et al., 2011; Cheng et al., 2014; Cornejo et al., 2014; Couceiro et al., 2013; Cowley et al., 2013; Cowley et al., 2014; Dudzinski et al., 2013; Evans et al., 2015; Froschauer et al., 2012; Froschauer et al., 2010; Hannig et al., 2012; Hauge & Riedel, 2012; Hou & Li, 2014; Kashibuchi & Sakamoto, 2001; Khanal et al., 2014; Klinge et al., 2009; LeRoy Heinrichs et al., 2008; Martin-Dorta et al., 2014; Mayer et al., 2013; Mortara et al., 2014; Tesei et al., 2014; Von Wangenheim et al., 2009; Wrzesien et al., 2010; Yin et al., 2012; Zhang et al., 2013; Zin & Yue, 2013)
Pre/Post/Post	This procedure is the same than *Pre/Post* procedure with an additionally stage. This new stage takes place after a period of weeks or months from the end of the second stage. The objective of this stage is to assess the level of retention of learned knowledge.	(Buttussi et al., 2013; Diehl et al., 2013; Quick et al., 2013; Wiezer et al., 2013)

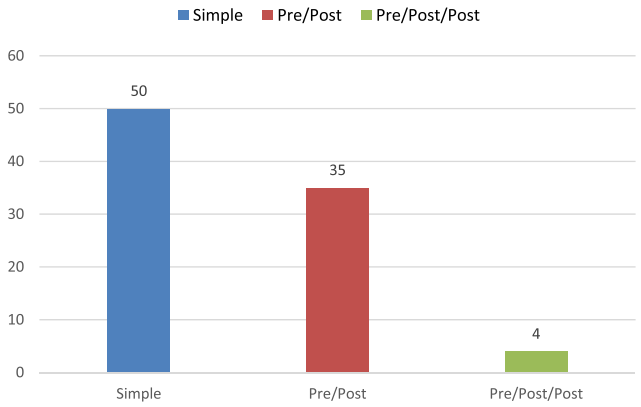

Fig. 14. Types of procedure.

(Cheng et al., 2014; Cook et al., 2012; Diehl et al., 2013; Froschauer et al., 2012; Hauge et al., 2013; Jarvis & Freitas, 2009; Kashibuchi & Sakamoto, 2001; Klinge et al., 2009; LeRoy Heinrichs et al., 2008; Quick et al., 2013; Rehm & Leichtenstern, 2012; Torrente et al., 2009; Wendel V. et al., 2013; Wiezer et al., 2013; Zin & Yue, 2013).

RQ6. What is the size of the population involved in the existing assessment experiences of the serious games?

After analyzing the methods, procedures and the quality characteristics that authors use to evaluate serious games, we explore the primary studies to find the size of the groups of people that take part in the evaluation of the serious games. In most of our primary studies, the average size of the population was less than 40 people, as Fig. 15 shows. There were a few studies using a population with a significant bigger size, such as 120 people or more, but this size was found to be either the size of the class in which the experiment took place or the number of students that participated in the experiments across several years, rather than a proper requirement of the evaluation method.

55% of the studies used a population with a size up to 40 people. A further analysis of these works revealed that a size between 11 up to 20 people is the most used size in the evaluation studies. Table 11 shows the primary studies classified attending at the ranges of population size used in the evaluations. Fig. 16 shows the distribution of studies according with the ranges of population size. From this analysis, we can conclude that the existing evaluations of serious games did not involve a large number of people.

From the information gathered in the analysis of our primary studies, it can be observed that only two of them are focused on the evaluation of a serious game proposal within the scope of the Software Engineering (Von Wangenheim et al., 2009; Von Wangenheim et al., 2012). From the analysis of both studies, we can conclude the following:

- They present and assess serious games to teach in the scope of Software Project Management.
- Their application domain is the higher education.
- Questionnaires are used as the technique for assessing both serious games.
- Both primary studies evaluate the learning outcomes of the serious games mainly as a function of: motivation, engagement and user experience.
- Both serious games follow different styles: one is a computer game, X-MED (Von Wangenheim et al., 2009), and another board game, Deliver! (Von Wangenheim et al., 2012), but the evaluation method and criteria are independent of the game platform or style.
- X-MED was evaluated by following a pre/post procedure, while Deliver! was evaluated by a simple procedure consisting on answering a questionnaire after playing the game.

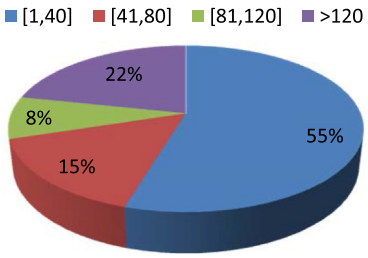

Fig. 15. Distribution of the population size.

Table 11
Primary studies within each population ranges.

Range	Primary studies
[1,10]	(Shah et al., 2014; Paliokas et al., 2011; Backlund P. et al., 2013; Backlund P. et al., 2011; Olsen et al., 2011; Manero et al., 2013; Hookham et al., 2015; Doukianou et al., 2014; Loreto et al., 2012; Terzidou et al., 2012; Burke et al., 2009)
[11,20]	(Vidani et al., 2010; Lopez-Basterretxea et al., 2014; Cornejo et al., 2014; Brown et al., 2007; Buttussi & Chittaro, 2010; Zhang et al., 2013; Adamo-Villani et al., 2013; Vaz de Carvalho et al., 2013; Lányi et al., 2012; Von Wangenheim et al., 2009; Birchall & Gatzidis, 2011; Birchall & Gatzidis, 2013; Cowley et al., 2014; Pohl et al., 2009; Lopez-Martinez et al., 2011; Froschauer et al., 2012; Jaligama & Liarokapis, 2011; Froschauer et al., 2010; Silva et al., 2013; Hess et al., 2014)
[21,30]	(Manera et al., 2015; Chan et al., 2012; Qin et al., 2010; Malliarakis et al., 2014; Wendel V., Gutjahr, Göbel, & Steinmetz, 2012; Hauge et al., 2013; Wrzesien et al., 2010; Von Wangenheim et al., 2012; Wendel V. et al., 2013; Rodríguez-Cerezo et al., 2014; Bernardini et al., 2014; Cederholm et al., 2011)
[31,40]	(Petrasova et al., 2010; Cook et al., 2012; Rehm & Leichtenstern, 2012; Creutzfeldt et al., 2013; Costantino et al., 2012; Yin et al., 2012; Pourabdollahian et al., 2012; Kerga et al., 2013; Gerling et al., 2011; Buttussi et al., 2013)
[41,80]	(Fuchslocher et al., 2011; Hannig et al., 2012; Enah et al., 2014; Ismailović et al., 2011; Cowley et al., 2013; Raybourn, 2009; Arnab et al., 2014; LeRoy Heinrichs et al., 2008; Hilborn et al., 2013; Mortara et al., 2014; Zin & Yue, 2013; Hou & Li, 2014; Basole et al., 2013; Dudzinski et al., 2013; Diehl et al., 2013)
[81,120]	(Mayer et al., 2013; Jarvis & Freitas, 2009; Dunwell et al., 2013; Wiezer et al., 2013; Torrente et al., 2009; Tesei et al., 2014; Couceiro et al., 2013; Cronan et al., 2012)
>120	(Cheng et al., 2014; Alamri et al., 2014; Klinge et al., 2009; Bascoul et al., 2013; Evans et al., 2015; Khanal et al., 2014; Berkovsky et al., 2010; Bevilacqua et al., 2014; Brom et al., 2010; Léger et al., 2012; Feldmann et al., 2014; Kashibuchi & Sakamoto, 2001; Rebolledo-Mendez et al., 2009; Pavlas et al., 2012; Ranchhod et al., 2012; Martin-Dorta et al., 2014; Muratet et al., 2011; Rughiniș, 2012; Qudrat-Ullah, 2010; Wildman & Reeves, 1996; Quick et al., 2013)

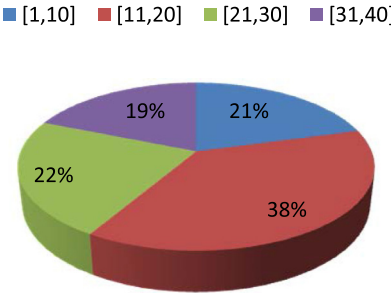

Fig. 16. Distribution of population size range [1, 40].

- Both evaluations use a population within the range from 1 to 40 people.
- The evaluation of both serious games concludes that the games cover the cognitive levels of remembering, understanding and application of the revised version of Bloom's taxonomy of educational objectives.
- The results of the evaluation of X-MED show that the participants consider the content and structure of the game appropriate, but it does not show any significant difference on learning effectiveness. Opposite to this, the results of the evaluation of Deliver! indicate the potential of such a game to contribute to the learning and the very positive effect of the game's application on the social interaction, engagement, immersion, attention and relevance to the course objectives.
- These two works take into account the revised version of Bloom's taxonomy of educational objectives.

6. The grey literature

In order to avoid the effect of publication bias in the published primary studies, it is important to search other sources of evidence, such as the grey literature (Thompson, 2001). The grey literature refers to print or electronic literature that is produced by government, academia, business and industry and is not controlled by commercial publishers (Auger, 1998). This can include materials such as unpublished studies, conference abstracts, conference proceedings, book chapters, government and agency reports, as well as unpublished doctoral dissertations. For that reason, in this section, we complement the searches in electronic databases with manual searches in the grey literature, in order to find studies that may be relevant to our research and were not found in the automatic searches.

To perform these manual searches, we applied the snowballing technique as follows:

1) **Step 1. Selection of primary studies.** Select the primary studies with the highest quality according to the quality assessment performed in Section 4.3.
2) **Step 2. Author snowballing.** Identify the authors of the selected primary studies and search their names in Google and Google Scholar to find publications that may be relevant to our research questions.

3) **Step 3. Backward snowballing.** Analyze the primary studies' list of references to find other relevant works not included in our list of primary studies.

First of all, we selected the best primary studies based on their score in the quality assessment. We selected the primary studies having a score of nine or more in their quality assessment. Step 1 concluded with a list of 13 studies selected following this criterion. Then, taking into account both the main authors and the co-authors of each publication, we identified fifty different authors (see Table 12).

Once we identified the authors, we analyzed their publications in the list resulting from the step 1 of this analysis. We considered that an author was relevant if they had authored more than one publication within the list of best primary studies. Only two of the identified authors (Cowley, B. and Ravaja, N.) fulfilled this criterion, since they had two different publications in the list of the highest quality primary studies (Cowley et al., 2013; Cowley et al., 2014). We manually searched "Benjamin Cowley" and "Niklas Ravaja" in Google and Google Scholar and did find other publications of these authors, but none of them in the scope of serious games evaluation. After analyzing the papers found, the reason we found for this is that the research area of these authors is not within the scope of serious games and their evaluation, being the primary studies we found the result of a single research effort aimed at developing and assessing a specific serious game. After finding this conclusion we decided to finish the second step of the manual searches.

The last step of the grey literature searches started with an analysis of the list of references of the primary studies in the list resulting from Step 1. We analyzed the title of each reference in order to find those works that may be related to serious games and their evaluation. In a first step, the papers cited in the selected works that included some of the following terms in the title, abstract or keywords were selected for further analysis: "serious games", "educational game", "simulation game", "evaluation", "assessment". We analyzed 569 references and found 21 different publications that had not been initially found in the electronic database searches. Table 13 shows the number of references analyzed for each paper in the list resulting from Step 1 and the number of relevant references found after the analysis.

After identifying the list of relevant publications, we searched them in both electronic databases, Google and Google Scholar, to get their full text and decide if any of them added new information to our study. 21 studies were found and analyzed (see Table A2). A summary of the main findings follows:

Table 12
Authors of the most quality primary studies.

Best primary studies	Author/s					
(Vaz de Carvalho, Lopes, & Gomes, 2013)	Vaz de Carvalho, C.		Lopes, M.P.		Gomes, D.F.	
(Couceiro, Papastergiou, Kordaki, & Veloso, 2013)	Rosana Margarida Couceiro		Marina Papastergiou		Maria Kordaki	Ana Isabel Veloso
(Ranchhod, Gurău, Loukis, & Trivedi, 2012)	Ranchhod, A.		Gur□u, C.		Loukis, E.	Trivedi, R.
(Cowley, Heikura, & Ravaja, 2013)	Cowley, B.		Heikura, T.		Ravaja, N.	
(Alamri, et al., 2014)	Alamri, A.	Hassan, M.M.	Hossain, M.A.	Al-Qurishi, M.	Aldukhayyil, Y.	Hossain, M.S.
(Rodríguez-Cerezo, Sarasa-Cabezuelo, Gómez-Albarrán, & Sierra, 2014)	Rodríguez-Cerezo, D.		Sarasa-Cabezuelo, A.		Gómez-Albarrán, M.	Sierra, J.L.
(Creutzfeldt, Hedman, Heinrichs, Youngblood, & Felländer-Tsai, 2013)	Creutzfel dt, J.	Hedman, L.	Heinrichs, L.	Youngblood, P.	Felländer-Tsai, L.	
(Von Wangenheim, Savi, & Borgatto, 2012)	Von Wangenheim, C.G.		Savi, R.		Borgatto, A.F.	
(Paliokas, Arapidis, & Mpimpitsos, 2011)	Paliokas, I.		Arapidis, C.		Mpimpitsos, M.	
(Vidani, Chittaro, & Carchietti, 2010)	Vidani, A.C.		Chittaro, L.		Carchietti, E.	
(Brom, Šisler, & Slavík, 2010)	Brom, C.		Šisler, V.		Slavik, R.	
(Martin-Dorta, et al., 2014)	Martin-Dorta, N.	Sanchez-Berriel, I.	Bravo, M.	Hernandez, J.	Saorin, J.L.	Contero, M.
(Cowley, Fantato, Jennett, Ruskov, & Ravaja, 2014)	Cowley, B.	Fantato, M.	Jennett, C.	Ruskov, M.	Ravaja, N.	

Table 13
Analysis of references.

Primary study	Number of references	Number of relevant references
(Vaz de Carvalho et al., 2013)	14	(Batista & Vaz de Carvalho, 2008; Gouveia, Lopes, & Vaz de Carvalho, 2011)
(Couceiro et al., 2013)	60	(Connolly, Stansfield, & McLellan, 2006; Kordaki, 2010; McGraw, Yoshimoto, & Seneff, 2009; Milone, Stegmayer, & Beber, 2009; Papastergiou, 2009; Sindre, Natvig, & Jahre, 2009; Virvou, Katsionis, & Manos, 2004)
(Ranchhod et al., 2012)	97	(Begum & Newman, 2009; Brennan, Willetts, & Vos, 2008; Chin, Dukes, & Gamson, 2009; Moizer, Lean, Mangles, & Warren, 2000; Pasin & Giroux, 2011; Towler, Lean, & Moizer, 2009)
(Cowley et al., 2013)	46	(Guillén-Nieto & Aleson-Carbonell, 2012)
(Alamri et al., 2014)	39	(Björk, Falk, Hansson, & Ljungstrand, 2001; Huang, 2011)
(Rodríguez-Cerezo et al., 2014)	47	(Eagle & Barnes, 2008)
(Creutzfeldt et al., 2013)	51	0
(Von Wangenheim et al., 2012)	46	(Navarro & Hoek, 2007)
(Paliokas et al., 2011)	36	0
(Vidani et al., 2010)	19	(Sliney & Murphy, 2008)
(Brom et al., 2010)	49	0
(Martin-Dorta et al., 2014)	28	0
(Cowley et al., 2014)	37	0
Total	569	21

- Most of the studies used:
  - Questionnaires as the main technique for serious games evaluation (Begum & Newman, 2009; Eagle & Barnes, 2008; Gouveia et al., 2011; Guillén-Nieto & Aleson-Carbonell, 2012; Huang, 2011; Kordaki, 2010; McGraw et al., 2009; Navarro & Hoek, 2007; Papastergiou, 2009; Sindre et al., 2009; Sliney & Murphy, 2008; Towler et al., 2009; Virvou et al., 2004).
  - Simple procedures to carry out the evaluation session of the serious games (Gouveia et al., 2011; Moizer et al., 2000; Towler et al., 2009; Sliney & Murphy, 2008) and Pre/Post (Begum & Newman, 2009; Eagle & Barnes, 2008; Guillén-Nieto & Aleson-Carbonell, 2012; Kordaki, 2010; McGraw et al., 2009; Papastergiou, 2009; Sindre et al., 2009; Virvou et al., 2004).
- Learning outcomes was the most assessed quality characteristics in order to observe the knowledge acquired by users (Begum & Newman, 2009; Guillén-Nieto & Aleson-Carbonell, 2012; Kordaki, 2010; McGraw et al., 2009; Moizer et al., 2000; Navarro & Hoek, 2007; Papastergiou, 2009; Sindre et al., 2009; Virvou et al., 2004).
- Education was the main domain of the evaluated serious games (Begum & Newman, 2009; Eagle & Barnes, 2008; Gouveia et al., 2011; Guillén-Nieto & Aleson-Carbonell, 2012; Huang, 2011; Kordaki, 2010; McGraw et al., 2009; Moizer et al., 2000; Navarro & Hoek, 2007; Papastergiou, 2009; Pasin & Giroux, 2011; Sindre et al., 2009; Sliney & Murphy, 2008; Towler et al., 2009; Virvou et al., 2004).
- Most of the serious games were computer based simulation games (Eagle & Barnes, 2008; Guillén-Nieto & Aleson-Carbonell, 2012; McGraw et al., 2009; Moizer et al., 2000; Navarro & Hoek, 2007; Papastergiou, 2009; Pasin & Giroux, 2011; Sliney & Murphy, 2008).
- The evaluation sessions were carried out with a population in a range from 4 to 144 people (Begum & Newman, 2009; Eagle & Barnes, 2008; Gouveia et al., 2011; Guillén-Nieto & Aleson-Carbonell, 2012; Huang, 2011; Kordaki, 2010; McGraw et al., 2009; Moizer et al., 2000; Navarro & Hoek, 2007; Papastergiou, 2009; Pasin & Giroux, 2011; Sindre et al., 2009; Sliney & Murphy, 2008; Towler et al., 2009; Virvou et al., 2004).
- Thanks to this analysis, one more study about the assessment of another serious game in the scope of software project management was identified. (Navarro & Hoek, 2007) presents and evaluate SimSE, a computer game based on software project simulation that supports several development methodologies and focuses on the development of abilities for software process management.

Given the information obtained from the grey literature analysis, we can conclude that this process did not provide any new findings regarding our research questions.

7. Threats to validity

The current systematic literature review has a number of factors that may have affected its validity. As with all reviews, it was limited by the search terms used, the journals included and the time period of papers published, because they limit the work performed. In order to overcome the influence of these limitations, we decided to apply the guidelines proposed by Kitchenham and Charters (2007) and, according to these guidelines, we defined the study protocol. In addition, we expanded the search process by conducting searches in the grey literature.

With respect the human resources, a single researcher selected the candidate studies. Given the number of publications found for the search engines, this could be a threat to the validity of the study because of possible human errors. Therefore, for the selection process to be as rigorous as possible, it was decided to perform a selection process in two stages, that

Table 14
Summary information.

RQ	Aspect	Most frequent answer
RQ1	Application domain	Education
RQ2	Type of game	Computer game
RQ3	Method	Questionnaire
RQ4	Quality characteristics	Learning outcomes, usability, user's experience
RQ5	Procedure	Simple
RQ6	Population	[1–40] people

dubious candidates could be assessed more deeply and be accepted or rejected in the second stage. The same limitation was in all the procedure of the review. This work could improve with involving more researchers in all the process of the review.

Finally, in order to assess the quality of each selected studies and whether these studies give us enough information to get conclusions regarding the defined research questions, we defined a Quality Assessment Questionnaire. For each question, we evaluated a YES or NO answer and we concluded that nearly 80% of the QAs were covered by YES values. This is a subjective method of evaluation, as it depends on the evaluator. But this process allowed us to add another step on the selection process according to the information achieved from each paper. The Quality Assessment step could be improved with the definition of more quality questions associated with each research question, so that they would provide a more realistic view on the quality of the selected papers.

8. Conclusions and future works

In this study, we identified 102 papers that evidence the procedures, techniques and methods used to assess serious games. We organized, classified and aggregated the information obtained to provide an answer to each of the six research questions set. The application domains in which the assessment experiences have taken place, the different types of serious games that have been assessed, the methods, procedures and techniques used, as well as the main features that have been assessed have been classified. These classifications provide a baseline for further research related to serious games assessment.

Considering the application domain in which more assessment experiences have taken place, we found that the educational domain, especially the higher education, is the domain where more studies evidenced assessment experiences, followed by health and wellness, and professional training and learning. We noticed that “computer games” is the type of game that attracts more interest towards assessment. Most of the studies analyzed selected the questionnaire as the main assessment method. The procedure followed in most of the experiences was a simple one, that is, without pre- and post-tasks, addresses to a population of less than 40 people. Most of the studies aim at the assessment of the educational effectiveness of the serious game. For this, the educational effectiveness is mostly defined as the learning outcomes, usability and user's experience.

Table 14 summarizes the most frequent answer to each aspect of the assessment of serious games found in this study.

The main objective of performing this systematic literature review was to find useful information about serious games assessment to allow us perform the assessment of our game. Taking the results obtained into account, we have selected to base out assessment method on the method followed by (Von Wangenheim et al., 2012). Although conceived to assess a board game, this method represents the main tendency summarized in Table 14. The method proposed assesses the quality features of motivation, user's experience and learning outcomes as the main factors of the game learning effectiveness. It is also a well-defined method based on Kirkpatrick's four levels of evaluation (Kirkpatrick & Kirkpatrick, 2006) and Bloom taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956), and the templates to develop the questionnaire for the assessment is also provided.

This study has helped us to analyze the evidence of the procedures, techniques and methods used to evaluate serious games to select a method to guide the assessment of our serious game. Our next steps are to finish the design of our assessment protocol and to carry out the evaluation sessions. In addition, we also plan to analyze and synthesize the information retrieved in this study regarding the different quality features that are measured when assessing serious games, so that a taxonomy of models for the evaluation of the different quality features can be proposed.

Acknowledgments

This work has been partially supported by the University of Cádiz and the Spanish Ministry of Science and Technology with ERDF funds under grants sol-201400048002-tra and TIN2013-46928-C3-2-R, respectively.

Appendix A

Table A1
Table of final primary studies.

Primary study	QA score
(Vaz de Carvalho et al., 2013)	9
(Couceiro et al., 2013)	9
(Ranchhod et al., 2012)	9
(Cowley et al., 2013)	9
(Alamri et al., 2014)	9
(Rodríguez-Cerezo et al., 2014)	9
(Creutzfeldt et al., 2013)	9
(Von Wangenheim et al., 2012)	9
(Paliokas et al., 2011)	9
(Vidani et al., 2010)	9
(Brom et al., 2010)	9
(Martin-Dorta et al., 2014)	9
(Cowley et al., 2014)	9
(Buttussi & Chittaro, 2010)	8
(Zhang et al., 2013)	8
(Pourabdollahian et al., 2012)	8
(Qudrat-Ullah, 2010)	8
(Von Wangenheim et al., 2009)	8
(Rughiniş, 2012)	8
(Mayer et al., 2013)	8
(Buttussi et al., 2013)	8
(Bascoul et al., 2013)	8
(Wiezer et al., 2013)	8
(Zin & Yue, 2013)	8
(Basole et al., 2013)	8
(Birchall & Gatzidis, 2013)	8
(Backlund P. et al., 2013)	8
(Diehl et al., 2013)	8
(Costantino et al., 2012)	8
(Hannig et al., 2012)	8
(Léger et al., 2012)	8
(Wendel V., Gutjahr, Göbel, & Steinmetz, 2012)	8
(Cook et al., 2012)	8
(Cronan et al., 2012)	8
(Yin et al., 2012)	8
(Pavlas et al., 2012)	8
(Rehm & Leichtenstern, 2012)	8
(Fuchslocher et al., 2011)	8
(Cederholm et al., 2011)	8
(Gerling et al., 2011)	8
(Muratet et al., 2011)	8
(Rebolledo-Mendez et al., 2009)	8
(Klinge et al., 2009)	8
(Kashibuchi & Sakamoto, 2001)	8
(Petrasova et al., 2010)	8
(Dunwell et al., 2013)	8
(Hou & Li, 2014)	8
(Wendel V. et al., 2013)	8
(Hess et al., 2014)	8
(Feldmann et al., 2014)	8
(Shah et al., 2014)	8
(Cheng et al., 2014)	8
(Dudzinski et al., 2013)	8
(Mortara et al., 2014)	8
(Arnab et al., 2014)	8
(Evans et al., 2015)	8
(Hookham et al., 2015)	8
(Bevilacqua et al., 2014)	8
(Doukianou et al., 2014)	8
(Malliarakis et al., 2014)	8
(Lopez-Martinez et al., 2011)	7
(Manero et al., 2013)	7
(Brown et al., 2007)	7
(Kerga et al., 2013)	7
(LeRoy Heinrichs et al., 2008)	7
(Hauge et al., 2013)	7
(Hilborn et al., 2013)	7
(Quick et al., 2013)	7
(Froschauer et al., 2012)	7

(*continued on next page*)

Table A1 (*continued*)

Primary study	QA score
(Loreto et al., 2012)	7
(Parsons et al., 2012)	7
(Lányi et al., 2012)	7
(Backlund P. et al., 2011)	7
(Birchall & Gatzidis, 2011)	7
(Jaligama & Liarokapis, 2011)	7
(Wrzesien et al., 2010)	7
(Froschauer et al., 2010)	7
(Berkovsky et al., 2010)	7
(Qin et al., 2010)	7
(Raybourn, 2009)	7
(Burke et al., 2009)	7
(Torrente et al., 2009)	7
(Bernardini et al., 2014)	7
(Adamo-Villani et al., 2013)	7
(Lopez-Basterretxea et al., 2014)	7
(Khanal et al., 2014)	7
(Cowan et al., 2010)	7
(Cornejo et al., 2014)	7
(Tesei et al., 2014)	7
(Manera et al., 2015)	7
(Hauge & Riedel, 2012)	6
(Chan et al., 2012)	6
(Terzidou et al., 2012)	6
(Ismailović et al., 2011)	6
(Jarvis & Freitas, 2009)	6
(Wildman & Reeves, 1996)	6
(Silva et al., 2013)	6
(El Borji & Khaldi, 2014)	6
(Enah et al., 2014)	6
(Olsen et al., 2011)	5
(Coelho et al., 2010)	5
(Pohl et al., 2009)	5

Table A2
Table of relevant references from the grey literature.

Relevant study	Primary study reference
(Gouveia et al., 2011)	(Vaz de Carvalho et al., 2013)
(Batista & Vaz de Carvalho, 2008)	
(Connolly et al., 2006)	(Couceiro et al., 2013)
(Kordaki, 2010)	
(McGraw et al., 2009)	
(Milone et al., 2009)	
(Papastergiou, 2009)	
(Sindre et al., 2009)	
(Virvou et al., 2004)	
(Begum & Newman, 2009)	(Ranchhod et al., 2012)
(Brennan et al., 2008)	
(Chin et al., 2009)	
(Moizer et al., 2000)	
(Pasin & Giroux, 2011)	
(Towler et al., 2009)	
(Guillén-Nieto & Aleson-Carbonell, 2012)	(Cowley et al., 2013)
(Björk et al., 2001)	(Alamri et al., 2014)
(Huang, 2011)	
(Eagle & Barnes, 2008)	(Rodríguez-Cerezo et al., 2014)
(Navarro & Hoek, 2007)	(Von Wangenheim et al., 2012)
(Sliney & Murphy, 2008)	(Vidani et al., 2010)

References

Abt, C. C. (2002). *Serious games*. Lanhan, MD: University Press of America.
ACM/IEEE-CS Joint Task Force on Computing Curricula. (2013). *Computer science curricula 2013*. ACM Press and IEEE Computer Society Press.
Adamo-Villani, N., Haley-Hermiz, T., & Cutler, R. (2013). Using a serious game approach to teach 'operator precedence' to introductory programming students. In *IV 2013-Proceedings of the 17th International Conference on Information Visualisation* (pp. 523–526). London, UK.

Alamri, A., Hassan, M. M., Hossain, M. A., Al-Qurishi, M., Aldukhayyil, Y., & Hossain, M. S. (January 2014). Evaluating the impact of a cloud-based serious game on obese people. *Computers in Human Behavior, 30*, 468–475.
Arnab, S., Perttula, A., & Suominen, M. (2014). Flow experience as a quality measure in evaluating physically activating serious games. *Lecture Notes in Computer Science, 8605*, 200–212.
Auger, P. (1998). *Information sources in grey literature* (4th ed.). London, UK: Bowker Saur.
Backlund, P., Taylor, A.-S. A., Engström, H., Johannesson, M., Lebram, M., Slijper, A., et al. (2011). Evaluation of usefulness of the Elinor Console for home-based stroke rehabilitation. In *VS-Games 2011-Proceedings of the 3rd International Conference on Games and Virtual Worlds for Serious Applications* (pp. 98–103). Athens, Greece.
Backlund, P., Taylor, A.-S. A., Engström, H., Johannesson, M., Lebram, M., Slijper, A., et al. (2013). Games on prescription! Evaluation of the Elinor Console for home-based stroke rehabilitation. *Lecture Notes in Computer Science, 7544*, 49–64.
Bascoul, G., Schmitt, J., Rasolofoarison, D., Chamberlain, L., & Lee, N. (2013). Using an experiential business game to stimulate sustainable thinking in marketing education. *Journal of Marketing Education, 35*(2), 168–180.
Basole, R. C., Bodner, D. A., & Rouse, W. B. (2013). Healthcare management through organizational simulation. *Decision Support Systems, 55*(2), 552–563.
Batista, R., & Vaz de Carvalho, C. (2008). Work in progress - learning through role play games. In *FIE 2008-Proceedings of the 38th IEEE Annual Frontiers in Education Conference* (pp. 1–8).
Begum, M., & Newman, R. (2009). Evaluation of students' experiences of developing transferable skills and business skills using a business simulation game. In *FIE 2009-Proceedings of the 39th IEEE Frontiers in Education Conference* (pp. 1–6).
Berkovsky, S., Freyne, J., Coombe, M., Bhandari, D., & Baghaei, N. (2010). Physical activity motivating games: virtual rewards for real activity. In *CHI 2010-Proceedings of the 28th annual CHI Conference on Human Factors in Computing Systems* (pp. 243–252). Atlanta, GA, United States.
Bernardini, S., Porayska-Pomsta, K., & Smith, T. (April 2014). ECHOES: an intelligent serious game for fostering social communication in children with autism. *Information Sciences, 264*, 41–61.
Bevilacqua, M., Ciarapica, F. E., Mazzuto, G., & Paciarotti, C. (2014). "Cook &Teach": learning by playing. *Journal of Cleaner Production* (in press).
Birchall, J., & Gatzidis, C. (2011). Elemental: an insight into the development and evaluation of a secondary education chemistry game using XNA. In *VS-Games 2011-Proceedings of the 3rd International Conference on Games and Virtual Worlds for Serious Applications* (pp. 32–39). Athens, Greece.
Birchall, J., & Gatzidis, C. (2013). The periodic table of elements via an XNA-powered serious game. *Lecture Notes in Computer Science, 7544*, 1–28.
Björk, S., Falk, J., Hansson, R., & Ljungstrand, P. (2001). Pirates! Using the physical world as a game board. In *Proceedings of Interact* (pp. 9–13).
Bloom, B., Engelhart, M., Furst, E., Hill, D., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. In *Handbook I: Cognitive domain*. New York: David McKay Company.
Brennan, R., Willetts, R., & Vos, L. (2008). *Student experiences of the use of a marketing simulation game*. Retrieved February 10, 2015, from Middlesex University Business School Discussion Paper http://nectar.northampton.ac.uk/1315/1/Brennan20081315.pdf.
Brereton, P., Kitchenham, B., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software, 80*(4), 571–583.
Brom, C., Šisler, V., & Slavík, R. (2010). Implementing digital game-based learning in schools: augmented learning environment of 'Europe 2045'. *Multimedia Systems, 16*(1), 23–41.
Brown, D., Shopland, N., Baftersby, S., Lewis, J., & Evett, L. (2007). Can serious games engage the disengaged?. In *Proceedings of the European Conference on Game-Based Learning* (pp. 37–46). Paisley, Scotland.
Burke, J. W., McNeill, M., Charles, D., Morrow, P., Crosbie, J., & McDonough, S. (2009). Serious games for upper limb rehabilitation following stroke. In *VS-Games 2009-Proceedings of the International Conference in Games and Virtual Worlds for Serious Applications* (pp. 103–110). Coventry, UK.
Buttussi, F., & Chittaro, L. (2010). Smarter phones for healthier lifestyles: an adaptive fitness game. *IEEE Pervasive Computing, 9*(4), 51–57.
Buttussi, F., Pellis, T., Cabas Vidani, A., Pausler, D., Carchietti, E., & Chittaro, L. (2013). Evaluation of a 3D serious game for advanced life support retraining. *International Journal of Medical Informatics, 82*(9), 798–809.
Calderón, A., & Ruiz, M. (2014). Bringing real-life practice in software project management training through a simulation-based serious game. In *CSEDU 2014 – Proceedings of the 6th International Conference on Computer Supported Education* (pp. 117–124). Barcelona, Spain.
Cederholm, H., Hilborn, O., Lindley, C., Sennersten, C., & Eriksson, J. (2011). The aiming game: using a game with biofeedback for training in emotion regulation. In *DiGRA 2011-Proceedings of the 5th International Conference on Digital Research Association: Think Design Play. Utrecht, Netherlands*.
Chan, W.-Y., Qin, J., Chui, Y.-P., & Heng, P.-A. (2012). A serious game for learning ultrasound-guided needle placement skills. *IEEE Transactions on Information Technology in Biomedicine, 16*(6), 1032–1042.
Cheng, M.-T., Su, T., Huang, W.-Y., & Chen, J.-H. (2014). An educational game for learning human immunology: what do students learn and how do they perceive? *British Journal of Educational Technology, 45*(5), 820–833.
Chin, J., Dukes, R., & Gamson, W. (2009). Assessment in simulation and gaming: a review of the last 40 years. *Simulation & Gaming, 40*(4), 553–568.
Coelho, T. B., Wesselius, M., & Papakonstantinou, C. (2010). OutRandom: addressing an underrated skill. In *Proceedings of the 3rd International Conference on Fun and Games* (pp. 166–170). Leuven, Belgium.
Connolly, T. M., Boyle, E. a., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*(2), 661–686.
Connolly, T. M., Stansfield, M., & McLellan, E. (2006). Using an online games-based learning approach to teach database design concepts. *Electronic Journal of e-Learning, 4*(1), 103–110.
Cook, N. F., McAloon, T., O'Neill, P., & Beggs, R. (2012). Impact of a web based interactive simulation game (PULSE) on nursing students' experience and performance in life support training - a pilot study. *Nurse Education Today, 32*(6), 714–720.
Cornejo, R., Hernandez, D., Tentori, M., & Favela, J. (2014). An ambient casual game to promote socialization and active ageing. *Lecture Notes in Computer Science, 8658*, 75–88.
Costantino, F., Di Gravio, G., Shaban, A., & Tronci, M. (2012). A simulation based game approach for teaching operations management topics. In *WSC 2012-Proceedings of the 2012 Winter Simulation Conference* (pp. 1–12). Berlin, Germany.
Couceiro, R. M., Papastergiou, M., Kordaki, M., & Veloso, A. I. (2013). Design and evaluation of a computer game for the learning of Information and Communication Technologies (ICT) concepts by physical education and sport science students. *Education and Information Technologies, 18*(3), 531–554.
Cowan, B., Sabri, H., Kapralos, B., Porte, M., Backstein, D., Cristancho, S., et al. (2010). A serious game for total knee arthroplasty procedure, education and training. *Journal of Cyber Therapy and Rehabilitation, 3*(3), 285–298.
Cowley, B., Fantato, M., Jennett, C., Ruskov, M., & Ravaja, N. (2014). Learning when serious: psychophysiological evaluation of a technology-enhanced learning game. *Educational Technology and Society, 17*(1), 3–16.
Cowley, B., Heikura, T., & Ravaja, N. (2013). Learning loops - interactions between guided reflection and experience-based learning in a serious game activity. *Journal of Computer Assisted Learning, 29*(4), 348–370.
Creutzfeldt, J., Hedman, L., Heinrichs, L., Youngblood, P., & Felländer-Tsai, L. (2013). Cardiopulmonary resuscitation training in high school using avatars in virtual worlds: an international feasibility study. *Journal of Medical Internet Research, 15*(1), e9.
Cronan, T. P., Léger, P., Robert, J., Babin, G., & Charland, P. (2012). Comparing objective measures and perceptions of cognitive learning in an ERP simulation game: a research note. *Simulation & Gaming, 43*(4), 461–480.
Diehl, L. A., Souza, R. M., Alves, J. B., Gordan, P. A., Esteves, R. Z., Jorge, M. L., et al. (2013). InsuOnline, a serious game to teach insulin therapy to primary care physicians: design of the game and a randomized controlled trial for educational validation. *Journal of Medical Internet Research, 2*(1), e5.
Doukianou, S., Petridis, P., Dunwell, I., & Cooper, J. (2014). A usability evaluation of game-based approaches assessing risk and delayed gratification. In *VS-GAMES 2014-Proceedings of the 6th International Conference on Games and Virtual Worlds for Serious Applications*.

Dudzinski, M., Greenhill, D., Kayyali, R., & Nabhani, S. (2013). The design and evaluation of a multiplayer serious game for pharmacy students. In *ECGBL 2013-Proceedings of the 7TH European conference on games based learning* (pp. 140–148).
Dunwell, I., Lameras, P., Star, K., de Freitas, S., Hendrix, M., & Arnab, S. (2013). Metycoon: a game-based approach to career guidance. In *VS-Games 2013-Proceedings of the 5th International Conference on Games and Virtual Worlds for Serious Applications* (pp. 1–6). Bournemouth, Dorset, UK.
Eagle, M. J., & Barnes, T. (2008). Wu's Castle: teaching arrays and loops in a game. In *Proceedings of the 13th Annual Conference on Innovation and Technology in Computer Science Education* (pp. 245–249). New York, NY, USA.
El Borji, Y., & Khaldi, M. (2014). Comparative study to develop a tool for the quality assessment of serious games intended to be used in education. *International Journal of Emerging Technologies in Learning (iJET), 9*(9), 50–55.
Enah, C., Piper, K., & Moneyham, L. (2014). Qualitative evaluation of the relevance and acceptability of a web-based HIV prevention game for rural adolescents. *Journal of Pediatric Nursing, 30*(2), 321–328.
Evans, K. H., Daines, W., Tsui, J., Strehlow, M., Maggio, P., & Shieh, L. (2015). Septris: a novel, mobile, online, simulation game that improves sepsis recognition and management. *Academic Medicine: Journal of the Association of American Medical Colleges, 90*(2), 180–184.
Feldmann, N., Adam, M., & Bauer, M. (2014). Using serious games for idea assessment in service innovation. In *ECIS 2014-Proceedings of the 22nd European Conference on Information Systems* (pp. 1–17). Tel Aviv, Israel.
Froschauer, J., Arends, M., Goldfarb, D., & Merkl, D. (2012). A serious heritage game for art history: design and evaluation of ThIATRO. In *VSMM 2012-Proceedings of the18th International Conference on Virtual Systems and Multimedia: Virtual Systems in the Information Society* (pp. 283–290). Milan, Italy.
Froschauer, J., Seidel, I., Gartner, M., Berger, H., & Merkl, D. (2010). Design and evaluation of a serious game for immersive cultural training. In *VSMM 2010-Proceedings of the 16th International Conference on Virtual Systems and Multimedia* (pp. 253–260). Seoul, South Korea.
Fuchslocher, A., Gerling, K., Masuch, M., & Kramer, N. (2011). Evaluating social games for kids and teenagers diagnosed with cancer. In *SeGAH 2011-Proceedings of the 1st International Conference on Serious Games and Applications for Health* (pp. 1–4). Braga, Portugal.
García-Mundo, L., Vargas-Enríquez, J., Genero, M., & Piattini, M. (2014). Contribuye el Uso de Juegos serios a Mejorar el Aprendizaje en el Área de la Informática. In *JENUI 2014-Proceedings of the XX Jornadas sobre la Enseñanza Universitaria de la Informática* (pp. 303–310). Oviedo, Spain.
Gerling, K., Fuchslocher, A., Schmidt, R., Kraemer, N., & Masuch, M. (2011). Designing and evaluating casual health games for children and teenagers with Cancer. *Lecture Notes in Computer Science, 6972*, 198–209.
Gouveia, D., Lopes, D., & Vaz de Carvalho, C. (2011). Serious gaming for experiential learning. In *FIE 2011-Proceedings of Frontiers in Education Conference* (pp. 1–6).
Guillén-Nieto, V., & Aleson-Carbonell, M. (2012). Serious games and learning effectiveness: the case of It's a Deal! *Computers & Education, 58*(1), 435–448.
Hannig, A., Kuth, N., Özman, M., Jonas, S., & Spreckelsen, C. (2012). eMedOffice: a web-based collaborative serious game for teaching optimal design of a medical practice. *BMC Medical Education, 12*(1).
Hauge, J. B., Duin, H., & Thoben, K.-D. (2013). The evaluation of serious games supporting creativity through student labs. *Lecture Notes in Computer Science, 8101*, 188–199.
Hauge, J. B., & Riedel, J. C. (2012). Evaluation of simulation games for teaching engineering and manufacturing. *Procedia Computer Science, 15*, 210–220.
Hess, M., Wiemeyer, J., Hamacher, K., & Goesele, M. (2014). Serious games for solving protein sequence alignments - combining citizen science and gaming. *Lecture Notes in Computer Science, 8395*, 175–185.
Hilborn, O., Cederholm, H., Eriksson, J., & Lindley, C. (2013). A biofeedback game for training arousal regulation during a stressful task: the space investor. *Lecture Notes in Computer Science, 8008*, 403–410.
Hookham, G., Nesbitt, K., Cooper, J., Croft, H., & Rasiah, R. (2015). Gamification for education: designing a pharmacy education game. *Lecture Notes in Computer Science, 8955*, 157–165.
Hou, H.-T., & Li, M.-C. (2014). Evaluating multiple aspects of a digital educational problem-solving-based adventure game. *Computers in Human Behavior, 30*, 29–38.
Huang, W.-H. (2011). Evaluating learners' motivational and cognitive processing in an online game-based learning environment. *Computers in Human Behavior, 27*(2), 694–704.
Ibrahim, I. (2011). Teaching project management for it students: methods and approach. In *2nd International Conference on Education and Management Technology* (pp. 185–191).
International Organization for Standardization (ISO). (2010). *ISO/IEC FDIS 25010: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*.
Ismailović, D., Pagano, D., & Brügge, B. (2011). Wemakewords - an adaptive and collaborative serious game for literacy acquisition. In *IADIS 2011-Proceedings of the International Conference Game and Entertainment Technologies* (pp. 3–10). Rome, Italy.
Jaligama, V., & Liarokapis, F. (2011). An online virtual learning environment for higher education. In *VS-Games 2011-Proceedings of the 3rd International Conference on Games and Virtual Worlds for Serious Applications* (pp. 207–214). Athens, Greece.
Jarvis, S., & Freitas, S. d (2009). Evaluation of an immersive learning programme to support triage training. In *VS-Games 2009-Proceedings of the International Conference in Games and Virtual Worlds for Serious Applications* (pp. 117–122).
Kashibuchi, M., & Sakamoto, A. (2001). The educational effectiveness of a simulation/game in sex education. *Simulation & Gaming, 32*(3), 331–343.
Kerga, E., Akaberi, A., Taisch, M., Rossi, M., & Terzi, S. (2013). Lean product development: serious game and evaluation of the learning outcomes. *IFIP Advances in Information and Communication Technology, 397*, 590–597.
Khanal, P., Vankipuram, A., Ashby, A., Vankipuram, M., Gupta, A., Drumm-Gurnee, D., et al. (October 2014). Collaborative virtual reality based advanced cardiac life support training simulator using virtual reality principles. *Journal of Biomedical Informatics, 51*, 49–59.
Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels*. San Francisco, CA: Berrett-Koehler Publishers.
Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*. Keele University and Durham University Joint Report.
Klinge, K., Rohmann, A., & Piontkowski, U. (2009). Intercultural sensitization with synthetic cultures: evaluation of a computer-based multimedia learning tool. *International Journal of Intercultural Relations, 33*(6), 507–515.
Kordaki, M. (2010). A computer card game for the learning of basic aspects of the binary system in primary education: design and pilot evaluation. *Education and Information Technologies, 16*(4), 395–421.
Lányi, C. S., Brown, D. J., Standen, P., & Butkute, V. (2012). Results of user interface evaluation of serious games for students with intellectual disability. *Acta Polytechnica Hungarica, 9*(1), 225–245.
Léger, P.-M., Cronan, P., Charland, P., Pellerin, R., Babin, G., & Robert, J. (2012). Authentic OM problem solving in an ERP context. *International Journal of Operations & Production Management, 32*(12), 1375–1394.
LeRoy Heinrichs, W., Youngblood, P., Harter, P. M., & Dev, P. (2008). Simulation for team training and assessment: case studies of online training with virtual worlds. *World Journal of Surgery, 32*(2), 161–170.
Lopez-Basterretxea, A., Mendez-Zorrilla, A., & Garcia-Zapirain, B. (2014). A telemonitoring tool based on serious games addressing money management skills for people with intellectual disability. *International Journal of Environmental Research and Public Health, 11*(3), 2361–2380.
Lopez-Martinez, A., Santiago-Ramajo, S., Caracuel, A., Valls-Serrano, C., Hornos, M. J., & Rodriguez-Fortiz, M. J. (2011). Game of gifts purchase: computer-based training of executive functions for the elderly. In *SeGAH 2011-Proceedings of the 1st International Conference on Serious Games and Applications for Health* (pp. 1–8). Braga, Portugal.
Loreto, I. D., Mora, S., & Divitini, M. (2012). Don't panic: enhancing soft skills for civil protection workers. *Lecture Notes in Computer Science, 7528*, 1–12.
MacDonell, S., Shepperd, M., Kitchenham, B., & Mendes, E. (2010). How reliable are systematic reviews in empirical software engineering? *IEEE Transactions on Software Engineering, 36*(5), 676–687.

Malliarakis, C., Satratzemi, M., & Xinogalos, S. (2014). CMX: Implementing an MMORPG for learning programming. *Proceedings of the European Conference on Games-based Learning, 1*, 346–355.
Manera, V., Petit, P.-D., Derreumaux, A., Orvieto, I., Romagnoli, M., Lyttle, G., et al. (2015). 'Kitchen and cooking,' a serious game for mild cognitive impairment and Alzheimer's disease: a pilot study. *Frontiers in Aging Neuroscience, 7*.
Manero, B., Fernandez-Vara, C., & Fernandez-Manjon, B. (2013). E-Learning takes the stage: from La Dama. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje, 8*(4), 197–204.
Marshall, G. (1998). *Snowballing technique – A dictionary of sociology*. Retrieved February 9, 2015, from Encyclopedia.com http://www.encyclopedia.com/doc/1O88-snowballingtechnique.html.
Martin-Dorta, N., Sanchez-Berriel, I., Bravo, M., Hernandez, J., Saorin, J. L., & Contero, M. (2014). Virtual blocks: a serious game for spatial ability improvement on mobile devices. *Multimedia Tools and Applications, 73*(3), 1575–1595.
Mayer, I., Wolff, A., & Wenzler, I. (2013). Learning efficacy of the 'hazard recognition' serious game: a quasi-experimental study. In *SGDA 2013-Proceedings of the 4th International Conference on Serious Games Development and Applications* (pp. 118–129). Trondheim, Norway.
McGraw, I., Yoshimoto, B., & Seneff, S. (2009). Speech-enabled card games for incidental vocabulary acquisition in a foreign language. *Speech Communication, 51*(10), 1006–1023.
Milone, D. H., Stegmayer, G., & Beber, D. (2009). Artificial life contest - a tool for informal teaching of artificial intelligence. In *CSEDU 2009-Proceedings of the First International Conference on Computer Supported Education* (pp. 119–124).
Moizer, J., Lean, J., Mangles, T., & Warren, M. (2000). Teaching introductory management using a computer simulation game. In *Proceedings of the 1st International Conference on Systems Thinking in Management* (pp. 456–461). Geelong, Australia.
Mortara, M., Catalano, C. E., Fiucci, G., & Derntl, M. (2014). Evaluating the effectiveness of serious games for cultural awareness: the Icura user study. *Lecture Notes in Computer Science, 8605*, 276–289.
Muratet, M., Torguet, P., Viallet, F., & Jessel, J. (2011). Experimental feedback on Prog&Play: a serious game for programming practice. *Computer Graphics Forum, 30*(1), 61–73.
Navarro, E. O., & Hoek, A. v (2007). Comprehensive evaluation of an educational software engineering simulation environment. In *CSEET 2007-Proceedings of the 20th Conference on Software Engineering Education & Training* (pp. 195–202).
Olsen, T., Procci, K., & Bowers, C. (2011). Serious games usability testing: how to ensure proper usability, playability, and effectiveness. *Lecture Notes in Computer Science, 6770*, 625–634.
Paliokas, I., Arapidis, C., & Mpimpitsos, M. (2011). PlayLOGO 3D: a 3D interactive video game for early programming education: let LOGO Be a game. In *VS-Games 2011-Proceedings of the 3rd International Conference on Games and Virtual Worlds for Serious Applications* (pp. 24–31). Athens, Greece.
Papastergiou, M. (2009). Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. *Computers & Education, 52*(1), 1–12.
Parsons, D., Petrova, K., & Ryu, H. (2012). Mobile gaming – a serious business! *WMUTE 2012-Proceedings of the IEEE 7th International Conference on Wireless, Mobile and Ubiquitous Technology in Education*, 17–24. Takamatsu, Kagawa, Japan.
Pasin, F., & Giroux, H. (2011). The impact of a simulation game on operations management education. *Computers & Education, 57*(1), 1240–1254.
Pavlas, D., Jentsch, F., Salas, E., Fiore, S. M., & Sims, V. (2012). The play experience scale: development and validation of a measure of play. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 54*(2), 214–225.
Petrasova, A., Cza, S., Chalmers, A., Farrer, J., & Wolke, D. (2010). The playability evaluation of virtual baby feeding application. In *VS-Games 2010-Proceedings of the 2nd International Conference on Games and Virtual Worlds for Serious Applications* (pp. 95–100). Braga, Portugal.
Pohl, M., Rester, M., & Judmaier, P. (2009). Interactive game based learning : advantages and disadvantages. *Lecture Notes in Computer Science, 5616*, 92–101.
Pourabdollahian, B., Taisch, M., & Kerga, E. (2012). Serious games in manufacturing education: evaluation of learners' engagement. In *VS-Games 2012-Proceedings of the 4th International Conference on Games and Virtual Worlds for Serious Applications* (pp. 256–265). Genoa, Italy.
Qin, J., Chui, Y.-P., Pang, W.-M., Choi, K.-S., & Heng, P.-A. (2010). Learning blood management in orthopedic surgery through gameplay. *IEEE Computer Graphics and Applications, 30*(2), 45–57.
Qudrat-Ullah, H. (2010). Perceptions of the effectiveness of system dynamics-based interactive learning environments: an empirical study. *Computers & Education, 55*(3), 1277–1286.
Quick, V., Corda, K., Chamberlin, B., Schaffner, D., & Byrd-Bredbenner, C. (2013). Ninja kitchen to the rescue: evaluation of a food safety education game for middle school youth. *British Food Journal, 115*(5), 686–699.
Ranchhod, A., Gurău, C., Loukis, E., & Trivedi, R. (April 2012). Evaluating the educational effectiveness of simulation games: a value generation model. *Information Sciences, 264*, 75–90.
Raybourn, E. M. (2009). Intercultural competence game that fosters metacognitive agility and reflection. *Lecture Notes in Computer Science, 5621*, 603–612.
Rebolledo-Mendez, G., Avramides, K., de Freitas, S., & Memarzia, K. (2009). Societal impact of a serious game on raising public awareness: the case of FloodSim. In *Sandbox 2009-Proceedings of the 2009 ACM SIGGRAPH Symposium on Video Games* (pp. 15–22). New Orleans, LA, United States.
Rehm, M., & Leichtenstern, K. (2012). Gesture-based mobile training of intercultural behavior. *Multimedia Systems, 18*(1), 33–51.
Rodríguez-Cerezo, D., Sarasa-Cabezuelo, A., Gómez-Albarrán, M., & Sierra, J.-L. (February 2014). 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.
Rughiniş, R. (2012). Serious games as input versus modulation: different evaluations of utility. In *Proceedings of the 26th Annual British Computer Society Interaction Specialist group Conference on People and Computers* (pp. 175–184).
Shah, N., Amirabdollahian, F., & Basteris, A. (2014). Designing motivational games for stroke rehabilitation. In *HSI 2014-Proceedings of the 7th International Conference on Human System Interactions* (pp. 166–171).
Shuttleworth, M. (2010). *Scientific control group*. Retrieved February 9, 2015, from Explorable.com https://explorable.com/scientific-control-group.
Silva, J. F., Silva, J. F., Rossetti, R. J., & Leca Coelho, A. (2013). A serious game for EVAcuation training. In *SeGAH 2013-Proceedings of the IEEE 2nd International Conference on Serious Games and Applications for Health* (pp. 1–6). Vilamoura, Portugal.
Sindre, G., Natvig, L., & Jahre, M. (2009). Experimental validation of the learning effect for a pedagogical game on computer fundamentals. *IEEE Transactions on Education, 52*(1), 10–18.
Sliney, A., & Murphy, D. (2008). JDoc: a serious game for medical learning. In *Proceedings of the 1st International Conference on Advances in Computer-Human Interaction* (pp. 131–136).
Terzidou, T., Tsiatsos, T., Dae, A., Samaras, O., & Chasanidou, A. (2012). Utilizing virtual worlds for game based learning: grafica, a 3D educational game in second life. In *ICALT 2012-Proceedings of the IEEE 12th International Conference on Advanced Learning Technologies* (pp. 624–628). Rome, Italy.
Tesei, A., Barbieri, A., Roceanu, I., & Beligan, D. (2014). Evaluation of "cultural awareness – Afghanistan Pre-deployment": a user study. *Lecture Notes in Computer Science, 8605*, 265–275.
Thompson, L. (2001). Grey literature in engineering. *Science and Technology Libraries, 19*(3), 57–73.
Torrente, J., Moreno-Ger, P., Fernández-Manjón, B., & Del Blanco, Á. (2009). Game-like simulations for online adaptive learning : a case study. *Lecture Notes in Computer Science, 5670*, 162–173.
Towler, M., Lean, J., & Moizer, J. (2009). An exploration of student perception of a business simulation game. *International Journal of Management Education, 7*, 69–79.
Vargas, J. A., García-Mundo, L., Genero, M., & Piattini, M. (2014). A systematic mapping study on serious game quality. In *EASE 2014-Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. London, UK*.
Vaz de Carvalho, C., Lopes, M., & Gomes, D. (2013). Serious games for lean manufacturing: the 5S game. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje, 8*(4), 191–196.

Vidani, A. C., Chittaro, L., & Carchietti, E. (2010). Assessing nurses' acceptance of a serious game for emergency medical services. In *VS-Games 2010-Proceedings of the 2nd International Conference on Games and Virtual Worlds for Serious Applications* (pp. 101–108). Braga, Portugal.

Virvou, M., Katsionis, G., & Manos, K. (2004). Combining software games with education: evaluation of its educational effectiveness. *Educational Technology & Society, 8*(2), 54–65.

Von Wangenheim, C. G., Savi, R., & Borgatto, A. F. (2012). DELIVER! – an educational game for teaching earned value management in computing courses. *Information and Software Technology, 54*(3), 286–298.

Von Wangenheim, C. G., Thiry, M., & Kochanski, D. (2009). Empirical evaluation of an educational game on software measurement. *Empirical Software Engineering, 14*(4), 418–452.

Wendel, V., Gutjahr, M., Battenberg, P., Ness, R., Fahnenschreiber, S., Limited, C., et al. (2013). Designing a collaborative serious game for team building using minecraft. In *ECGBL 2013-Proceedings of the 7th European Conference on Games Based Learning* (pp. 569–578).

Wendel, V., Gutjahr, M., Göbel, S., & Steinmetz, R. (2012). Designing collaborative multiplayer serious games for collaborative learning: escape from Wilson Island - a multiplayer 3D serious game for collaborative learning in teams. In *CSEDU 2012-Proceedings of the 4th International Conference on Computer Supported Education* (pp. 199–210).

Wiezer, N., Roozeboom, M. B., & Oprins, E. (2013). Serious gaming used as management intervention to prevent work-related stress and raise work-engagement among workers. In *DHM 2013-Proceedings of the 4th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management* (pp. 149–158). Las Vegas, NV, United states.

Wildman, S., & Reeves, M. (1996). The utilization and evaluation of a simulation game in pre-registration nurse education. *Nurse Education Today, 16*(5), 334–339.

Wrzesien, M., López, D., & Raya, M. (2010). Learning ecology issues of the mediterranean sea in a virtual aquatic world - pilot study. *Journal of Cyber Therapy and Rehabilitation, 3*(3), 255–260.

Yin, L., Ring, L., & Bickmore, T. (2012). Using an interactive visual novel to promote patient empowerment through engagement. In *FDG 2012-Proceedings of the International Conference on the Foundations of Digital Games* (pp. 41–48). Raleigh, NC, United States.

Zhang, J., Caldwell, E. R., & Smith, E. (2013). Learning the concept of Java inheritance in a game. In *CGAMES 2013-Proceedings of the 18th International Conference on Computer Games: AI, Animation, Mobile, Inte3ractive Multimedia, Educational & Serious Games* (pp. 212–216). Louisville, KY.

Zin, N. A., & Yue, W. S. (2013). Design and evaluation of history digital game based learning (DGBL) software. *Journal of Next Generation Information Technology, 4*(4), 9–24.

Zyda, M. (September 2005). From visual simulation to virtual reality to games. *Computer, 38*, 25–32.