A Systematic Mapping Study on Serious Game Quality

Juan A. Vargas
Instituto Tecnológico de
Cd. Victoria, Mexico
JuanAntonio.Vargas@alu.uclm.es

Lilia García-Mundo
Instituto Tecnológico de
Cd. Victoria, Mexico
LiliaCarmen.Garcia@alu.uclm.es

Marcela Genero
Institute of Technologies and Information Systems,
University of Castilla-La Mancha
Ciudad Real, Spain
Marcela.Genero@uclm.es

Mario Piattini
Institute of Technologies and Information Systems,
University of Castilla-La Mancha
Ciudad Real, Spain
Mario.Piattini@uclm.es

ABSTRACT

Context: A Serious Game (SG) is a game for purposes other than entertainment [12]. SGs are currently in widespread use and their popularity has begun to steadily increase; their application areas now extend not only to education, but also to military, health and corporate [9] [12] sectors. SGs are of vital importance at present, as they can be a means to achieve relevant goals from both a personal and an institutional point of view. This may take place in fields as diverse as defense, education, scientific exploration, health care, emergency management, city planning, engineering, religion, and politics. The number of users of these systems grows each day, signifying that their impact is very high, and it is precisely for this reason that more extensive research on SG quality is needed.
Objective: The aim of this study is to discover the current state of SG quality initiatives, identifying gaps that merit future investigation.
Method: We conducted a systematic mapping study (SMS) on SG quality, following the guidelines proposed by Kitchenham and Charters [7]. We selected 112 papers found in six digital libraries until April of 2013.
Results: Since 2007, research on SG quality proves to have grown very significantly. Research has focused mostly on addressing the effectiveness of SGs (78.57%), in addition to several entertainment characteristics that are principally related to pleasure (62.50%). The most widely-researched software artifact was the final product (97.32%), with design coming very far behind (7.14%). Less than half of all the research reviewed had been validated by means of experiments, and in most of these cases, experiments were conducted by the same researchers who had proposed the SG. The majority of experiments have not been replicated. The most common research outcome was questionnaires, closely followed by the confirmation of knowledge. Most of these outcomes evaluated the quality of a particular SG.
Conclusion: Results show that SG quality has undergone a very important growth, thus making SG quality an area of opportunity for future research. Researchers are mainly concerned with demonstrating or confirming the effectiveness of SGs, but very little research has been conducted as regards the characteristics of playability that make SGs more effective. Since effectiveness and playability are evaluated in the final product there is a need to provide quality assurance methods that incorporate quality issues from the early stages of SG development. Further empirical validation is also needed, and in particular, external replications must be performed in order to corroborate and generalize the findings obtained.

Categories and Subject Descriptors
D.2.0 [**Software Engineering**]: General – standards.
D.2.9 [**Software Engineering**]: Management – Software quality assurance (SQA).

General Terms
Measurement, Standardization.

Keywords
Serious Games, Quality, ISO 25010, Systematic Mapping Study.

1. INTRODUCTION
Although SGs have commonly been defined as "games in which education is the primary goal, rather than entertainment" [9], this definition does not fully define what a SG is. According to Ben Sawyer [9], the word 'serious' refers to the purpose of the game and not to the content of the game itself. SGs are not only intended for education; serious purpose takes in a wide range of application areas [12]. A broader definition of SG provided by Michel Zyda is "a mental contest, played with a computer in accordance with specific rules that uses entertainment to further government or corporate training, education, health, public policy, and strategic communication objectives" [16]. In a more general and simplified way, it can be said that SGs are games for purposes other than entertainment [12].
SGs are a fast-emerging area of opportunity, in addition to being a rapidly growing market [9]. SGs are currently in widespread use; their popularity has begun to steadily increase and their application areas now extend not only to education but also to military, health and corporate [9][12] sectors. In 2012, worldwide revenues for game-based learning (a type of SG) alone amounted to 1.5 billion dollars. With a global growth rate of 8 % a year, it is forecasted that by 2017 worldwide revenue will reach 2.3 billion dollars [1].

Although one might initially think that a game cannot be serious, play is an important part of the learning process. Van Eck has pointed out that play is a primary socialization and learning mechanism common to all human cultures and many animal species [13]. Connolly et al [3], for their part, claim that SGs provide features which, according to modern learning theories, make learning more effective.

Researchers agree that the widespread acceptance of SGs is owing to both their positive impact and the advantages they have over traditional learning methods [16][12][15]. Amongst the benefits of SGs we can indicate the following : 1) SGs allow learners to experience situations that it would otherwise be impossible to experience in real life owing to aspects related to costs, resources, time, security, etc. [12]. 2) There is evidence that SGs support the acquisition of knowledge, that they are more effective than traditional instructional methods as regards training cognitive skills and that they have a promising use in the learning of fine-grid motor skills that require excellent hand-eye coordination [15]. 3) SGs enable the employment potential of staff to be enhanced, while simultaneously improving their technical capabilities. They also make it possible to catch up with and keep abreast of technological development; they foster local development and strengthen regional cohesion [8]. 4) In the particular case of Advergames, longer information retention and more effective memorization is achieved, signifying that more complex messages can be delivered without boring a captive audience; these people will thus remember the brand specifications better.

Although research has been conducted on several topics related to SGs, more extensive research is needed on SG quality. SGs are critically important at present, as they can be a means to achieve relevant goals, from both a personal and an institutional point of view. They may be used in fields as diverse as defense, education, scientific exploration, health care, emergency management, city planning, engineering, religion, and politics. In addition, the number of users of these systems grows each day, signifying that their social impact is very high. It is for this reason that SG quality is so critical; they are not just another variety of software (in which it is already assumed that quality is important). They can have a major impact on many areas of society and on many users, and it is therefore our duty as researchers and computer professionals to ensure their quality.

In order to discover the current state of SG quality initiatives and to identify gaps that merit future rigorous investigation, we decided to conduct a systematic mapping study (SMS) following the guidelines proposed by Kitchenham and Charters [7]. The main goal of this paper is to summarize all the tasks developed in the planning and execution of the SMS.

The remainder of the paper is organized as follows: Section 2 presents the related work. In Section 3 the planning of the SMS is described, and in Section 4 we explain how the SMS was conducted. Section 5 sets out the data synthesis and results, and finally, in Section 6 the conclusions are presented, along with our future work.

2. RELATED WORK

We found six literature reviews on SGs published in the last ten years, which are summarized as follows.

- Kirriemuir and McFalane [6] conducted a literature review focused on the design of games for both entertainment and learning for school-age children. They proposed 3 research questions: (1) What happens during the game-playing process?, (2) Can conventional computer games be used as a vehicle for formal learning and (3) What components of conventional computer games can be used to learn about software or in practice?. They concluded that a better understanding of the potential of these games is needed, that the game development industry needs to understand the constraints, resources, and the requirements of educational games, and that there are as yet a small number of games that have a clear contribution to make to the educational agenda.
- Hays [4] performed a literature review on educational games, focusing on the empirical research as regards the instructional effectiveness of games. The review included 105 papers published up to 2005. 48 of them included empirical data regarding the effectiveness of educational games. Some of the conclusions resulting from the review are: The empirical research on the effectiveness of educational games includes research on different tasks, age groups, and types of games. Although research has shown the effectiveness of some games on learning, this should not be generalized to all games in all learning areas for all learners. Educational games should be embedded in instructional programs that give the learners the opportunity to reflect on their experience with the game and understand how this experience supports the instructional objectives. It should also provide feedback on how close the learner is to achieving the goal.
- Susi et al [12] published a report which reviewed several papers concerning SGs - principally the advantages of SGs and the positive and negative effects of SGs on learning. They concluded that although SGs are generally considered to increase various skills, there may be a lack of evidence to support this claim. What is more, it would appear that there is no conclusive answer to the question of evidence as regards the supposed benefits and potential consequences of games and game play.
- Wouters et al [15] conducted a review of the literature focusing on empirical evidence regarding the learning outcomes of SGs. The review was carried out in the summer of 2008 and included papers from the previous 10 years. 28 papers containing empirical data on the effectiveness of SGs were found. These authors concluded that SGs potentially improve the acquisition of knowledge and cognitive skills, and that they seem to be promising for the acquisition of fine-grid motor skills and the accomplishment of attitudinal change. However, not all game features increase the effectiveness of the games.
- Connolly et al [3] carried out a systematic literature review that gathered empirical evidence on the positive impacts and outcomes of computer games and SGs with regard to learning and engagement. The review focused primarily on UK research in which the participants were over 14 years old. The review included 129 papers published between January 2004 and February 2009. Their findings revealed that the most frequently occurring outcomes and impacts were knowledge acquisition/content understanding and affective and motivational outcomes. They also concluded that although empirical evidence concerning the effectiveness of game-based learning was found, there is a need for more rigorous evidence of their effectiveness.
- In a research report, McClarty et al [10] presented an overview of the theoretical and empirical evidence behind five key claims about the use of digital games in education. The claims were that digital games (1) were built on sound

learning principles, (2) provided more engagement for the learner, (3) provided personalized learning opportunities, (4) taught 21st century skills, and (5) provided an environment for authentic and relevant assessment. The review included 87 papers published from 1996 to 2011. They concluded that digital games can facilitate learning, but that it is difficult to draw stronger conclusions about the educational impact of digital games.

The literature review of Kirriemuir and McFalane [6] is the only one not to focus on the effectiveness of SGs but rather on the design, and does not describe a systematic selection process. The review of Hays [4] focuses on the empirical evidence as regards the effectiveness of SGs. Susi et al [12] focus on empirical research related to the advantages of SGs, along with the positive and negative effects in learning, and do not describe a systematic selection. Wouters et al [15] focus on the empirical research dealing with the learning effectiveness of SGs, but the review is carried out on a fairly small scale and only 28 papers are considered. Connolly et al [3] focus on empirical research concerning the learning effectiveness and engagement of games and SGs, but the scope of the search is predominantly restricted to one country (UK) and to one age group in particular. McClarty et al [10] focus on the theoretical and empirical evidence regarding the advantages and benefits of educational games, but do not describe a systematic selection. In a nutshell, 5 of the 6 literature reviews presented above have focused on the learning effectiveness or positive effects of SGs [3,4,10,12,15], of which 4 have focused only on empirical research [3,4,10,15]. The most recent literature review is that of McClarty et al [10], while the others were carried out at least 5 years before ours. With regard to the procedure followed to perform the literature review, in most cases there is no information about the number of articles reviewed, the selection procedure etc.

The literature review presented in this paper is different from the previous one in several respects:

- Goal. We collect the existing literature on SG quality, and do not only focus on empirical evidence.
- Period of time. The period of time covered is longer.
- Procedure. This literature review has been carried out in a systematic and rigorous manner, following the guidelines provided by Kitchenham and Charters [7]. These guidelines aim to present a fair evaluation of a research topic by using a trustworthy, rigorous, and auditable methodology.

3. PLANNING THE REVIEW

The planning is related to developing the protocol, which establishes a controlled procedure with which to conduct the review. Our protocol includes objectives, research questions, a search strategy, selection strategy (inclusion/exclusion criteria), study selection procedure, data extraction strategy, and data synthesis.

The aim of this SMS is to determine the current state of SG quality initiatives and to identify gaps that merit further research, in order to provide practitioners and researchers with relevant and updated information. Based on this objective, five research questions are proposed. The research questions and their motivation are described in Table 1.

Table 1. Research questions

Research questions	Main motivation
RQ1. What particular quality characteristics of SGs have been investigated by researchers?	To identify the quality characteristics of SGs that have been addressed by researchers, and map them onto the quality characteristics proposed in the ISO/IEC 25010 [5]
RQ2. What is the nature of the research on SG quality?	To discover what proposals have been produced by the research work on SG quality.
RQ3. What research methods have been used to investigate SG quality?	To determine whether or not the research has been validated and to discover which research method was used to validate it.
RQ4. What software artifacts has the research on SG quality been focused on?	To discover whether SG quality has been researched throughout the whole software development lifecycle or whether it has focused solely on certain software artifacts.
RQ5. What have the application areas of SGs in the research of SG quality been?	To identify the SG application areas in which quality has been investigated.

The search string was constructed using the steps described in (Brereton 2007).

- Derive major terms from the questions.
- Identify alternative spellings, synonyms and related terms for major terms.
- Use the Boolean OR to incorporate alternative spellings, synonyms and related terms.
- Use the Boolean AND to link the major terms.

We initially carried out a pilot search using "Serious Game" AND "Quality" as our search string. As this search returned very few useful results, we decided not to include the term "Quality" in the search string because it is too ambiguous. We then included the terms "evaluat*", "assess"*, "measur*" and "test*" because we found that these terms were frequently used in papers dealing with the quality of SGs. The major search terms are “Evaluation” and “Serious Game”. The alternative spellings, synonyms and terms related to the major terms are shown in Table 2.

Table 2. Major search terms and their alternative terms

Major Terms	Alternative terms
Evaluation	(evaluat* OR assess* OR measur* OR test*)
Serious game	(“serious game” OR “educational game” OR “learning game” OR “educational computer game”)

The intention of this SMS was to discover all papers that present any research related to SG quality (with or without empirical data), that are written in English and have been published until April 2013. The start of the publication period was not established because we wished to discover since when SG quality proposals have existed. Papers were excluded according to the selection criteria shown in Table 3.

The selection procedure of the studies was executed with the search string defined by the first two authors of this paper. The selection process of the studies was conducted in two stages. In

the first stage, the selection of the studies was performed by reviewing the title, the abstract and the keywords of the studies; only those papers that dealt with SG quality were selected. Based on the set of papers selected in the first stage, the second stage consisted of reading the full texts of these papers and applying the inclusion and exclusion criteria. During the entire procedure, the mechanics were: one of the authors reviewed the paper, and the other author then verified it. Any discrepancies were resolved by a consensus between the four authors, taking into account the full text of the paper.

Table 3. Inclusion and exclusion criteria

Inclusion criteria	• Papers that fulfill the search string. • Journals, conferences and workshop papers. • Papers written in English. • Papers published until April 2013 (inclusive).
Exclusion criteria	• Papers not focusing on SG quality. • Papers available only in the form of abstracts or PowerPoint presentations. • Papers that presented an abstract of a workshop submission. • Duplicate papers (same research in different databases). • Papers in which SG quality is mentioned only as a general introductory term, or where there is no proposal related to quality among the paper´s contributions.

The searches were performed in electronic collections that contain a wide variety of computer science journals: SCOPUS database, Science@Direct with the subject Computer Science, Wiley InterScience with the subject of Computer Science, IEEE Digital Library, ACM Digital Library, and the SPRINGER database. A summary of the search strategy used is shown in Table 4.

Table 4. Search strategy

Databases Searched	• Scopus • Science@Direct (subject Computer Science) • Wiley InterScience (subject Computer Science) • IEEE Digital Library • ACM Digital Library • Springer database
Target items	• Journal papers • Workshop papers • Conference papers
Search applied to	• Title • Abstract • Keywords
Language	Papers written in English.
Publication period	Until April 2013 (inclusive)

The activity of classifying the data gathered was facilitated with the use of a two-part form. The first part was related to the metadata of the paper (paper ID, extractor name, reviewer name, and title), while the second part contained the multidimensional classification scheme. A set of five dimensions were used to classify the research, based on the research questions described above. Each of the five dimensions consisted of several categories. The Quality Characteristics categories were defined on the basis of the ISO-25010 standard [5]. The scheme suggested by [14] was used as a starting point to determine the categories for the research method. The categories for the remaining dimensions were defined prior to the extraction of information from the SMS and were refined on the basis of the data extracted. These dimensions and their categories are summarized in Table 5; a detailed description of the classification scheme can be accessed at http://alarcos.esi.uclm.es/SMS-SeriousGamesQuality/.

A quantitative synthesis method will be used to present the results of this SMS. The quantitative synthesis is based on counting the primary studies that will be classified according to

Table 5. Summary of the classification scheme

Dimensions	Categories
Quality characteristic	*Quality in use model:* Effectiveness, efficiency, satisfaction (usefulness, trust, pleasure) *Product quality model:* functional suitability (functional completeness, functional correctness, functional appropriateness) performance efficiency (time behavior) usability (appropriateness recognizability, learnability, user error protection, operability, user interface aesthetics) reliability (fault tolerance) portability (adaptability)
Research result	Questionnaire, knowledge, scale, guide, tool, heuristic, framework, method
Research method	Proposal, evaluation, validation, philosophical, opinion or personal experience
Software artifact	Requirement, design, code, final product
Application area	Skills development, awareness, health, training, general application, education (Computer, Languages, History, Mathematics, Physics, Natural Sciences, Geography, etc.)

the dimensions and categories defined (see Table 5); the combinations of the dimensions and categories will be displayed using bubble plots.

4. CONDUCTING THE REVIEW

The SMS process was completed in twelve months; this period included the time needed for planning, conducting and reporting. Four researchers took part in the whole process. The outline of the SMS is shown in Table 6. The planning of the SMS began in December 2012. All papers related to SG quality published until April 2013 were retrieved in May 2013. We found 1236 papers; after the title and abstract of each paper had been reviewed, the number of papers selected was reduced to 262 (excluding papers not related to SG quality). We excluded 93 duplicate papers (the same paper in a different database) and proceeded to extract and review the full texts of the remaining 169 papers. Inclusion and exclusion criteria were applied (to the full text), and 34 more papers were discarded. We excluded 23 more duplicate papers (the same study published more than once). This analysis was used to refine the extraction and classification schemes, identify

primary studies, eliminate follow-up studies and make final classifications. The final 112 papers were then analyzed and the results were interpreted (Figure 1). The list of 112 primary studies is available at http://alarcos.esi.uclm.es/SMS-SeriousGamesQuality/.

The protocol was defined by the two first authors of this work; and was then iteratively reviewed and refined by the last two authors. The identification and selection of primary studies was performed by the two first authors. In order to reduce the risk of a publication being incorrectly included in or excluded from the SMS, each paper was reviewed by at least two authors. In those cases in which these two authors had conflicting views, a third and a fourth author were required to review the publication and make a final decision.

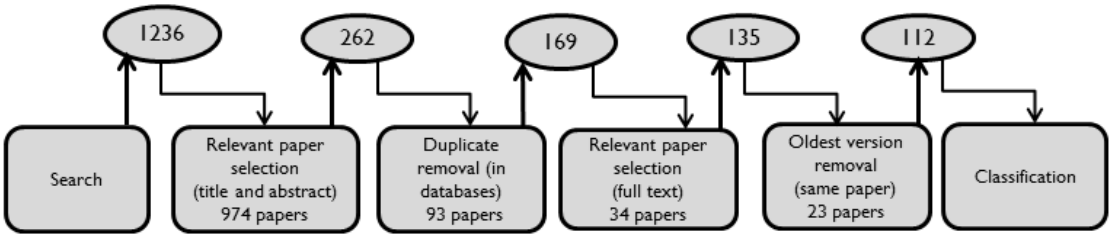

Figure 1. Selection process

Table 6. Outline of the SMS

Chronology	Step	Activities	Outcome
December 2012	**Planning**	Protocol development	Review protocol
April 2013	**Conducting**	Data retrieval Study selection (title and abstract) Removal of duplicates Extraction of files of the papers	Metadata information of 1236 papers Metadata information of 262 papers selected Metadata information of 169 papers selected Repository of papers (169 papers)
July 2013	**Planning**	Protocol improvement Pilot data extraction	Data extraction form (classification scheme refined), 169 papers reviewed
August 2013	**Conducting**	Study selection, classification (full text) Removal of oldest versions of same papers Data synthesis	Data extraction form complete, 135 papers classified 112 papers reviewed and classified
November 2013	**Reporting**	Report on the stages and activities undertaken during the development of the SMS	Final report

Some search engines in databases or digital libraries have limitations when using complex Boolean search strings. When a database or digital library did not allow the use of a complex Boolean search string, we designed different search strings for each database manually. The purpose was to obtain the same results that had been achieved using the original search string.

5. DATA SYNTHESIS AND RESULTS

In this section we present the answers to each of the research questions formulated, in addition to the results obtained by carrying out a thorough search of the findings of several of these research questions. We shall also discuss and interpret some of the results obtained.

5.1 Results by research question

The answers to each of the research questions formulated were obtained by synthesizing the data gathered from the papers selected. A summary of the quantitative results of the research questions from RQ2 to RQ5 is presented in Table 7.

5.1.1 RQ1. What particular quality characteristics of SGs have been investigated by researchers?

The process used to match the characteristics in the ISO/IEC 25010 standard [5] with the characteristics investigated in the paper is described as follows. We read the full text of the paper in search of quality characteristics that were addressed by the authors, and then looked at the standard for the characteristic or characteristics that in our opinion best matched the characteristics found in the paper. In the review of the full text of the selected papers, we found that in most of them the authors used several terms to refer to the quality characteristics being researched; in the majority of cases these terms did not match those specified in the standard. We also found that on a number of occasions there was no match between the characteristics that the authors were investigating and the characteristics in the standard. In other cases the characteristic investigated was equivalent to more than one characteristic in the standard. The table of 112 primary studies which contains this correspondence is available at http://alarcos.esi.uclm.es/SMS-SeriousGamesQuality/.

The results for RQ1 revealed that most of the papers selected addressed more than one quality characteristic or sub-characteristic. We found that the quality model most frequently investigated, in 88 papers, is the quality in use model, as shown in Figure 2. We also found that 43 of the articles researched a particular characteristic or sub-characteristic of the product quality model (Figure 3).

The characteristics most frequently addressed by the quality in use model were effectiveness (78.57%) and satisfaction (64.29%). Satisfaction, was mostly addressed by the sub characteristic pleasure (62.50%), and to a far lesser extent by the sub-

characteristic utility (13.39%) (Figure 2). The characteristics of the quality product model most frequently researched were, meanwhile, usability (45.54%) distantly followed by functional suitability (8.93%). We observed that usability was most frequently researched through the use of the operability sub-characteristic (38.39%), closely followed by the user interface aesthetics sub-characteristic (35.71%), and to much lesser degree by the learnability sub-characteristic (8.93%) (see Figure 3). These results could be explained by the fact that researchers are principally concerned with demonstrating or confirming that the SG meets the (serious) purpose for which it was created, which is why they research effectiveness and usefulness. But researchers are also interested in knowing whether the SG is capable of providing enjoyment and entertainment, which is the part of the SG as regards playability. That is why pleasure, the user interface aesthetics and operability are also the focus of research. We believe that this explains why the other characteristics of SG quality that are not related to these aspects, such as efficiency, performance efficiency or security, have been neglected. Similar findings were reported by Connolly et al [3], who found that the most frequently researched issue was the effectiveness of knowledge acquisition, but that many papers also reported enjoyment and engagement. We also found that very little research has been conducted on the relationship between the effectiveness of SGs and the characteristics of playability that make them effective. It seems that researchers are very interested in knowing whether SGs are effective but not in what makes them effective.

We believe that it would be interesting to investigate which playability aspects have an influence on SGs' effectiveness. Wouters et al [15] came to similar conclusions when stating that a better understanding of the underlying motivational processes such as enjoyment and engagement in SGs is required.

Table 7. Summary of the quantitative results from RQ2 to RQ5

Research question	Possible answers	Papers	Percentage
RQ2. Research result	questionnaire	43	38.39
	knowledge	37	33.04
	scale	3	2.68
	guide	2	1.79
	tool	2	1.79
	heuristic	6	5.36
	framework	13	11.60
	method	6	5.36
RQ3. Research method	**Empirical evidence:**	**41**	**36.61**
	Validation	39	34.82
	experiment	26	23.21
	quasi-experiment	13	11.61
	evaluation	2	1.79
	Non Empirical evidence:	**71**	**63.39**
	philosophical	12	10.71
	proposal	59	52.68
RQ4. Software artifact	requirement	2	1.79
	design	8	7.14
	code	2	1.79
	product	109	97.32
RQ5. Application area	education	68	60.71
	skills development	8	7.14
	awareness	1	0.89
	health	8	7.14
	training	13	11.62
	general application	4	12.50

5.1.2 RQ2. What is the nature of the research on SG quality?

This research question refers to a result generated in the research, i.e., the proposal that is made in the research addressing SG quality. The results showed that the most common output was the questionnaire (43 papers, 38.39%). This type of research produced questionnaires to illustrate the proposals presented. In a very close second place (37 papers, 33.04%), was the confirmation of knowledge. These studies evaluated the effectiveness of SGs, confirming whether participants improved their learning. In third place were the frameworks (13 papers, 11.60%), and this category contained articles that proposed frameworks, checklists for carrying out evaluations, etc. In fourth place were heuristics and methods (6 papers each, 5.36%), followed by scales, which were the result of 3 studies (2.68%). Finally, in seventh place, were guides and tools (2 paper each, 1.79%). Few proposals dealing with SG quality were presented as a framework (13), method (6) or heuristic (6).

Most of the proposals that evaluate the quality of SGs do so for a particular game, but in a different manner, and there is therefore no agreement among researchers as to how to evaluate the same quality characteristic. For example we found that effectiveness is evaluated by means of frameworks, methods, questionnaires, etc. This problem was also indicated by McClarty et al [10], who claimed that one of the reasons why the results of the effectiveness of SGs are not conclusive is that researchers have not agreed on either the definitions or the methodologies used for evaluation.

Only 28.57% (32 papers) produced an outcome that could be applied to any kind of SGs. There is a need for research into SG quality, whose outcome can be applied to any SG in general and that will enable researchers to validate any SG proposed in the same way.

Figure 2. Distribution of papers according to characteristics of the quality in use model

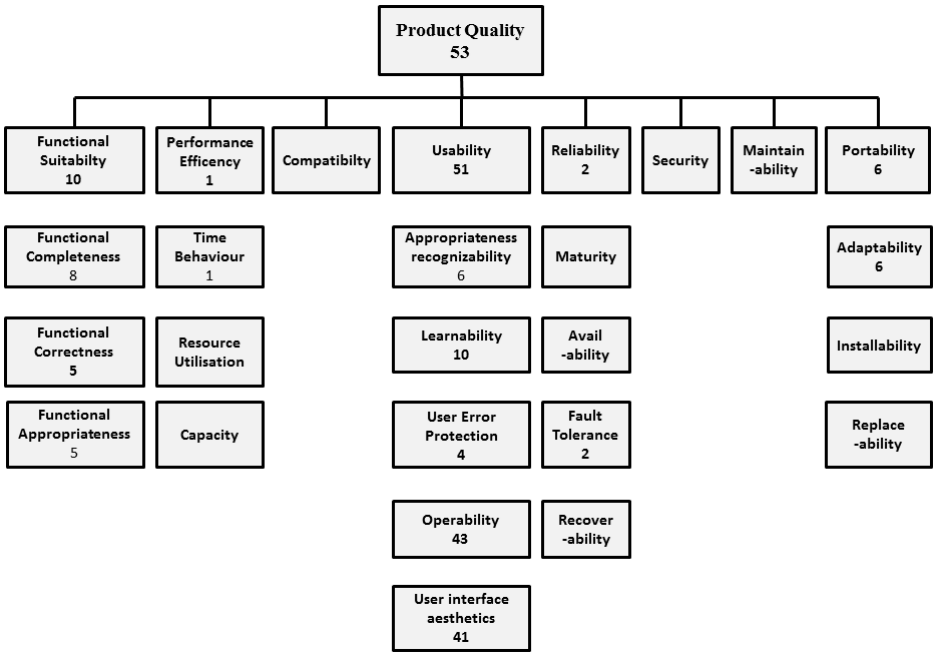

Figure 3. Distribution of papers according to characteristics of the product quality model

5.1.3 RQ3. What research methods have been used to investigate SG quality?

This question was answered by using the classification of research approaches proposed by Wieringa et al [14], as recommended in Petersen et al [11]. This classification makes it possible to classify empirical research into either validation or evaluation. The scheme also presents the classification of non-empirical research, which contains the categories of *proposal papers, philosophical papers, opinion papers* and *personal experience papers*. In this review we found only papers that were classified by the categories of *validation, evaluation, proposal* and *philosophical*; the majority of the papers were related to *proposal* and *philosophical* (non-empirical research) (63.39 %).

The results showed that *proposal* (52.68%) stood out as the dominant research method. The second most common research method used was *validation* (34.82%); in third place was *philosophica*l (10.71%), and finally in last place was *evaluation* (1.79%). *Experiment* (23.21%) was the validation method that was used most, followed by the *quasi-experiment* (11.61%). Less than half of all the research work reviewed in the papers selected had been validated (41, 36.61%); in these works the validation was done by conducting an experiment or a quasi-experiment. In most of these papers, empirical validation had been conducted by the same researchers who had proposed the SG, and in most cases the studies had been not replicated. We believe that there is a need to replicate empirical validations in order to corroborate and generalize the findings obtained.

5.1.4 RQ4. What software artifacts has the research on SG quality been focused on?

The results showed that of the 112 papers reviewed, 109 dealt with SG quality in the final product (97.32%). This kind of SG quality is produced after the product has been developed, or when a final version is ready. 8 papers (7.14%) dealt with SG quality in the design, and the request and the code were addressed by 2 papers each (1.79%). The results show that the evaluation of quality in current SG development practices is often deferred to late stages in the SG game development cycle, thus signifying that quality problems from early stages may be propagated to late stages of the development, consequently making their detection and correction a very expensive task. For example the introduction of pedagogical or playability aspects in the design stage might have an impact on the effectiveness of SGs. Addressing SG quality in the final product is necessary if we are to explore user behavior, but there is also a need to address the quality of SGs from the early stages of the development to ensure the quality of the final product.

5.1.5 RQ5. What have the application areas of SGs in the research of SG quality been?

This SMS revealed that education was by far the most widely-used application area in which SGs were referred to (60.71%). This type of research dealt with SGs that were intended to teach some specific topic. In second place were SGs that were implied in some kind of professional training (11.62%). Thirdly, with the same percentage, were health-related SGs, and SGs that were used for skill development (7.14%). The health-related SGs were those whose aim was to show how to perform a surgical procedure or those used by patients as physical therapy. Finally, in fourth place, were SGs for some kind of awareness (0.89%). Certain SGs (12.50%) could not be sorted into any of the above areas, and were therefore classified as general applications. Because of the large amount of articles that were related to education, this application area was sub-classified into sub-areas to which the SG teaching referred (Figure 4). In the order of this sub-classification was: firstly SGs for Computing (16), secondly, SGs related to Languages (9), thirdly, SGs with which to teach History (8). After them, in fourth place, were SGs for the teaching of Mathematics (6) and in fifth place, SGs about Physics (5), followed by SGs for Natural Science teaching (4). In last place were SGs related to teaching Geography (3). Given the diversity of the knowledge sub-areas in 17 papers, these were grouped into a classification called *Other*.

These results show the importance and wide acceptance that SGs have had in education, but simultaneously highlight the lack of research work into quality in other areas of SG application. They indicate an opportunity for research in application areas such as training, healthcare and skill development, etc.

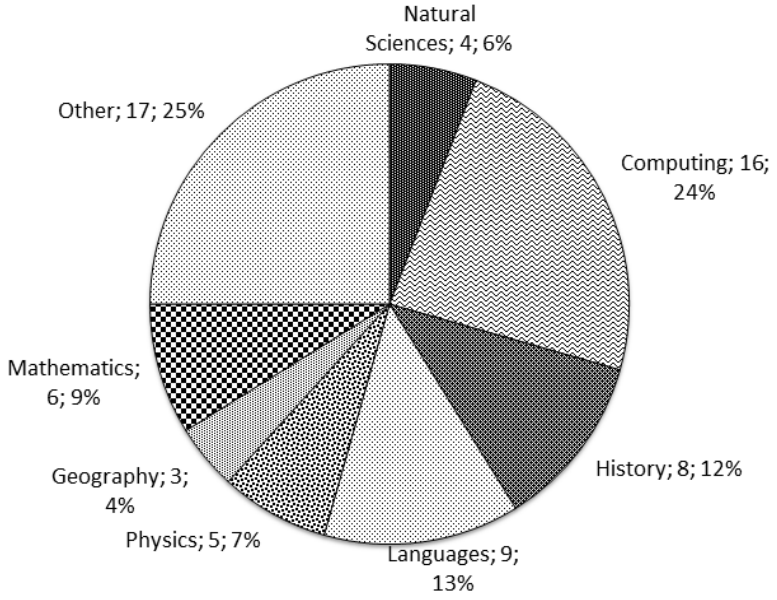

Figure 4. Distribution of papers per education area

5.2 Combination of results

Apart from the individual analysis of each research questions, the most was made of the results by combining the findings of several research questions.

5.2.1 Combining RQ1, RQ2, RQ3 and RQ4

Combining the results of research questions RQ1, RQ2, RQ3 and RQ4, as shown in Figure 5, we observe that the major outcomes of research are the questionnaire and confirmation of knowledge; in both cases SG quality is addressed in the final product. However, papers whose results are a confirmation of knowledge show empirical validation, while those with questionnaires do not. Effectiveness and pleasure are the most frequently researched characteristics, but papers that address effectiveness show more empirical validation than those addressing pleasure. These quality characteristics are addressed in the final product. In most cases, research on the effectiveness or the pleasure of SGs produce a questionnaire or a confirmation of knowledge as a result. When the outcome of the research is a questionnaire, the research method is always the proposal of a solution; this is presented with a proof of concept by means of a small example. When the outcome of the research is a confirmation of knowledge, this is usually validated through an experiment or quasi-experiment.

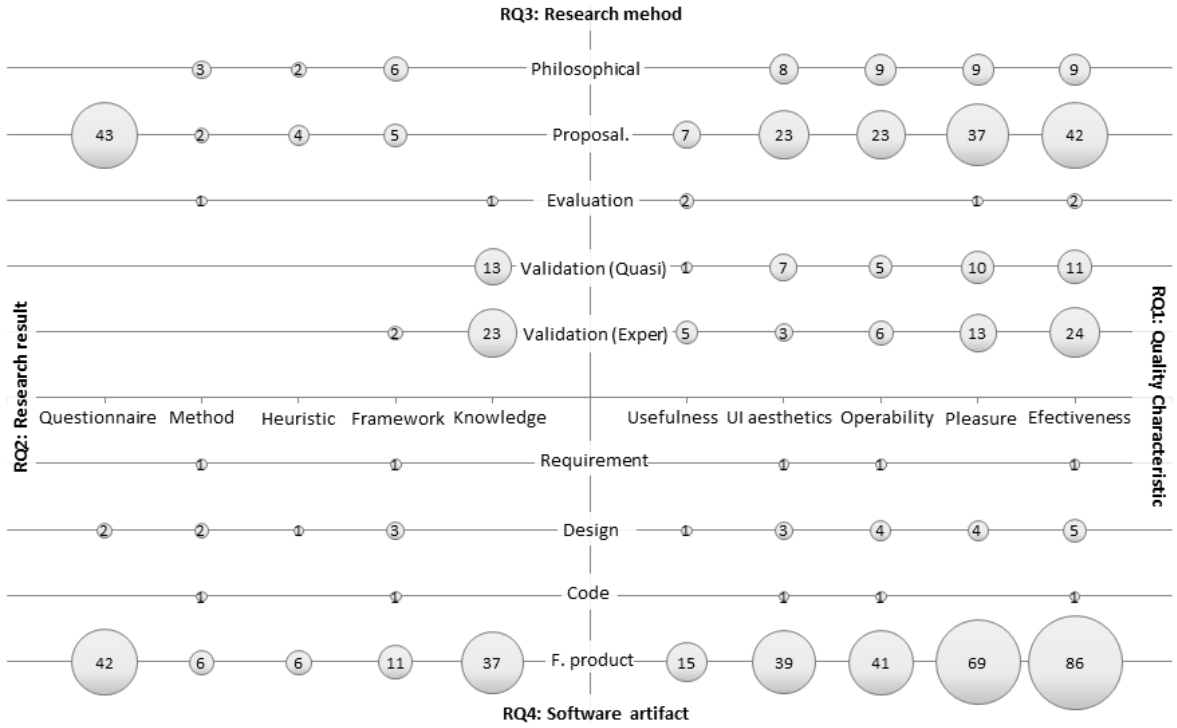

Figure 5. Combination of results of RQ1, RQ2, RQ3 and RQ4

5.3 Additional results

Starting in 2007, there is a clear progression in the number of publications focused on SG quality that appear each year, with the highest point being reached in 2012. The number of publications in 2013 was lower, because this study only considered publications until April 2013. Results show that since 2008, SGs have undergone a very significant growth, and have in recent years become a "hot topic", thus making SG quality an area of opportunity for future research.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an SMS related to SG quality; we selected 112 papers from the 1236 found in 6 digital libraries until April 2013. Results also show that SG quality has undergone a very important growth, rising from 3 papers in 2007 to 34 papers in 2012. In recent years SG quality has therefore become a "hot topic", thus making SG quality an area of opportunity for future research.

The results show that researchers are mainly concerned with demonstrating or confirming the effectiveness of SGs in addition to their capability of providing enjoyment and entertainment, but that very little research has been carried out as regards the characteristics of playability that make SGs more effective. Other characteristics of SGs have barely been addressed, such as efficiency, performance efficiency or security. Since effectiveness and playability are evaluated in the final product there is a need to provide quality assurance methods that incorporate quality issues from the early stages of the SG development focusing, for example, on quality characteristics that may have an impact on an SG's effectiveness, such as pedagogical and playability aspects introduced in the design stage. Approximately half of the proposals that deal with SGs quality have been empirically validated by means of experiments carried out by the same researchers who propose the SG, and they have not been replicated. Although 28.57% (32 papers) of the studies produced have an outcome that can be applied to any SG, only 2.68% (3 papers) of these outcomes have been validated.

Our interpretation of the review results has allowed us to identify some possible research opportunities:

- Propose, apply and validate a quality assurance method that incorporates quality evaluation from the early stages of the SG development cycle. It would preferably be possible to apply this method to any kind of SG.
- Investigate which playability aspects have an influence on SGs' effectiveness.
- Carry out more empirical validation. Both internal and external replications are needed in order to corroborate and generalize the findings obtained.

We also plan to carry out a Systematic Literature Review in order to synthesize the empirical evidence on SG quality.

7. ACKNOWLEDGMENTS

This research has been funded by the GEODAS-BC project (Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional FEDER, TIN2012-37493-C03-01).

We would also like to thank the Instituto Tecnológico de Ciudad Victoria and PROMEP for granting us the scholarship that made it possible to complete the research work presented in this paper.

8. REFERENCES

[1] Ambient Insight Research: The 2012-2017 Worldwide Game-based Learning and Simulation-based Markets, 2013.

Retrieved October 22, 2013, from Ambient Insight Research: http://www.ambientinsight.com/Resources/Documents/AmbientInsight_SeriousPlay2013_WW_GameBasedLearning_Market.pdf.

[2] Brereton, P., Kitchenham, B., Budgen, D., Turner, M., and Khalil, M. Lessons from Applying the Systematic Literature Review Process Within the Software Engineering Domain. *The Journal of Systems and Software, 80,* 571-583.

[3] Connolly, T., Boyle, E., MacArthur, E., Hainey, T., and Boyle, J. A Systematic Literature Review of Empirical Evidence on Computer Games and Serious Games. *Computers and Education, 59* (2), 661-686.

[4] Hays, R.T. *The Effectiveness of Instructional Games: A Literature Review and Discussion*. Technical Report 2005-004. Naval Air Warfare Center Training Systems Division, 2005.

[5] ISO. ISO/IEC FDIS 25010: Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - System and Software Quality Models. ISO (International Organization for Standardization), 2010.

[6] Kirriemuir, J., and McFarlane, A. *Literature Review in Games and Learning.* Report 8. Graduate School of Education of University of Bristol, 2004.

[7] Kitchenham, B., and Charters, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering.* Technical Report EBSE-2007-01. Software Engineering Group of Keele University Durham UK, 2007.

[8] LUDUS: How can one Benefit from Serious Games, 2010. Retrieved November 19, 2013, from LUDUS: http://www.ludus-project.eu/sgbenefits.html.

[9] Michael, D., and Chen, S. *Serious Games: Games That Educate, Train, and Inform.* Thomson Course Technology PTR, Boston Ma, 2006.

[10] McClarty, K.L., Orr, A., Frey, P.M., Dolan, R., Vassileva, V., and McVa, A. *A Literature Review of Gaming in Education.* Research Report. Pearson Education, 2012.

[11] Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. Systematic Mapping Studies in Software Engineering. In *Conference on Evaluation and Assessment in Software Engineering*, (Bari-Italy, 2008), British Computer Society, 68-77.

[12] Susi, T., Johannesson, M. and Backlund, P. *Serious Games – An Overview*. Technical Report HS- IKI -TR-07-001. School of Humanities and Informatics University of Skövde Sweden, 2007.

[13] Van Eck, R. Digital Game-Based Learning: It's Not Just the Digital Natives Who Are Restless. *EDUCAUSE Review, 41* (2), 17-30.

[14] Wieringa R., Maiden, N., Mead, N., and Rolland, C. Requirements Engineering Paper Classification and Evaluation Criteria: A Proposal and a Discussion. *Requirements Engineering, 11* (1), 102-107.

[15] Wouters, P., Van der Spek, E., and Van Oostendorp, H. Current Practices in Serious Game Research: A Review from a Learning Outcomes Perspective. In Connolly, T., Stansfield, M., and Boyle, L. *Games-Based Learning Advancements for Multi-Sensory Human Computer Interfaces: Techniques and Effective Practices*, IGI Global, Hershey PA USA, 2009, 232-250.

[16] Zyda, M. From Visual Simulation to Virtual Reality to Games. *Computer, 38* (9), 25-32.