ICGA Journal 41 (2019) 116–137
DOI 10.3233/ICG-190111
IOS Press

Let players evaluate serious games. Design and validation of the Serious Games Evaluation Scale

Emmanuel Fokides [a,*], Penelope Atsikpasi [a], Polyxeni Kaimara [b] and Ioannis Deliyannis [b]
[a] *Department of Primary Education, University of the Aegean*
[b] *Department of Audiovisual Arts, Ionian University*

Abstract. Although a number of researchers consider serious games as effective teaching/learning tools, the literature is fragmented when it comes to the factors that shape the users' experience and views. We are currently lacking a comprehensive tool that simultaneously examines their effectiveness while contrasting user views. The study is an attempt to fill this gap. It reports the development and validation of a scale which initially included seventy-two items belonging to thirteen factors. A total of 542 university students played two serious games and the aforementioned questionnaire was administered to them. The exploratory and confirmatory factor analysis revealed that twelve factors and fifty-three items should be retained. The final version of the Serious Games Evaluation Scale demonstrated satisfactory reliability and validity. The factorial structure of the questionnaire and its implications for research and practice are also discussed.

Keywords: Effectiveness, factors, scale, serious games, users' views

1. INTRODUCTION

One of the earliest, as well as one of the most widely used definitions for serious games (SGs), states that they are deliberately educational games/applications; their recreational aspects are minimal or even absent (Abt, 1970). While many consider SGs as effective for promoting learning (Connolly, Boyle, MacArthur, Hainey & Boyle, 2012; de Freitas, 2018; Erhel & Jamet, 2019; Lamb, Annetta, Firestone, & Etopio, 2018) in diverse educational scenarios and domains (Girard, Ecalle, & Magnan, 2013), a number of interconnected issues and concerns were not adequately addressed, until today. SGs' impact on knowledge acquisition and how this knowledge is transferred to real-life conditions remain rather unclear (Blumberg, Almonte, Anthony & Hashimoto, 2013). The design and development of SGs is a multifaceted interdisciplinary process involving experts from many fields. A typical team is made up of software engineers, usability experts, media designers, and education experts. In addition, other experts may join forces during development, depending on the type of game, the supporting environment, and the teaching/learning requirements of the system. The end result of this collaborative effort requires robust evaluation methods in order to assess each and every aspect of SGs. Yet, this task is difficult, given that there are many different genres of SGs covering an equally large number of learning subjects. Hence, a specific evaluation method for one SG cannot be easily generalized to all SGs (Ravyse, Blignaut, Leendertz, & Woolner, 2017). This methodological deficiency led researchers to believe that we lack an established methodology for measuring their effectiveness (e.g., Serrano-Laguna, Manero, Freire, & Fernández-Manjón, 2018). Indeed, research examining many salient factors that render SGs effective is rather scarce (Ravyse et al., 2017). What is more, a number of factors suffer from definitional problems (e.g., immersion and presence) or there

*Corresponding author. Tel.: +302241099238; E-mail: fokides@aegean.gr.

is no common consensus of what sub-features they encompass, causing some confusion on how they can be evaluated (Fokides, Atsikpasi, Kaimara, & Deliyannis, 2019).

Our approach raises the importance of user evaluation. We believe that it is important to identify how users view SGs as they can pinpoint the positive and negative system aspects, enabling designers to identify the important characteristics by analyzing the essence of the users' experience. Then again, there is no clear definition of the term "user experience", as there are conceptualization and as well as measurement problems (Buck, Khan, Fagan & Coman, 2018; Koeffel, Hochleitner, Leitner, Haller, Geven, & Tscheligi, 2010). For the industry, the "user experience" is viewed as a synonym of usability and user-centered design, while researchers primarily focus on subjective constructs such as emotions, feelings, and sensations (Buck et al., 2018). Even though this field has become a major concern for both researchers and practitioners in human-computer interaction (Lallemand, Gronier, & Koenig, 2015), its role in determining the success of educational applications and their acceptance by educators and learners is not fully recognized (O'Brien, 2016).

The study at hand presents a comprehensive tool that can measure as many as possible of the factors that come into play, in an attempt to fill the research gap described above. Thirteen factors and seventy-two items were included in an initial draft of a scale and a project was implemented in order to gather data that would allow the examination of its factorial structure and validity. The reasoning for selecting these factors, the research methodology, and the results of the exploratory and confirmatory factor analysis of the scale are discussed in the coming sections.

2. BACKGROUND-RESEARCH OBJECTIVE

As the learning outcomes are what interests most stakeholders, these were the focus of the majority of SGs assessments. In addition, an array of other factors was also examined, scattered across a large number of studies. For example, usability and engagement were considered by some (e.g., All, Castellar, & Van Looy, 2015). For others, engagement and motivation were the most significant factors (e.g., Huang, Huang, & Tschopp, 2010). Khan and Webster (2017) examined narration as a contributing factor. In an effort to holistically examine SGs, Steiner, Hollins, Kluijfhout, Dascalu, Nussbaumer, Albert, and Westera (2015) proposed the inclusion of learning effectiveness, usability, and enjoyment. Gameplay, challenge, fun, feedback, interaction, scenario, immersion, learning-game integration, and game design were emphasized by many (e.g., Marsh, 2011; Muratet, Viallet, Torguet, & Jessel, 2009). Calderón and Ruiz (2015) identified a total of eighteen features that are important, namely: game design and aesthetics, user's satisfaction, performance, usability-ease of use-playability-learnability, engagement, usefulness, understandability, motivation, educational aspects, learning outcomes, user's behavior-attitude-emotions, efficacy, social impact, enjoyment, acceptance, and interface. SGs encapsulate both leisure and "serious" purposes. Consequently, the users' experience and, in turn, the users' views for a given SG are influenced by its pedagogical as well as its gaming aspects.

The above studies used different factors, different genres of SGs were examined, and the learning subjects or the learning settings were also diverse. What is more, the literature regarding SGs' assessment can be classified as rather fragmented. Further investigation revealed that there is no common consensus on the definition of some factors. For example, the terms "presence" and "immersion" were used interchangeably and were even examined using identical or similarly worded questions. The same also holds true for "usability" and "ease of use." Evidently, more research is needed in order to establish which SGs' features are important in shaping their learning effectiveness and the users' views (Hersh & Leporini, 2018). Given that, what the study sought to accomplish was the development of

a tool that would allow the simultaneous examination of a broad range of SG's factors responsible for shaping the users' views. Toward this end, by probing more into the relevant literature, thirteen factors commonly used for assessing the users' views and experiences were brought to light:

- **Enjoyment.** The fun and enjoyment users feel when playing SGs is probably the most significant reason for using such applications. As expected, enjoyment is used in most SGs' evaluation frameworks (e.g., Steiner et al., 2015).
- **Motivation.** The motivational appeal is also one of the primary reasons for playing SGs (Garris, Ahlers, & Driskell, 2002). Although when playing SGs the users are expecting some extrinsic rewards (as this is done in an educational context), SGs provide intrinsic motivation; the activity per se is rewarding.
- **Presence-immersion.** These subjective experiences suffer from definitional problems (Fokides & Atsikpasi, 2018) as both are used for describing a range of similar feelings the players have during playing. Presence is a psychological state in which the virtual objects are perceived as real (Ivory & Kalyanaraman, 2007). Immersion, on the other hand, is the feeling of "being in the game"; the players lose track of time and of their surroundings (Ermi & Mäyrä, 2005). Given that many studies used these terms interchangeably, in the present study the term "presence" summarizes both.
- **Playability.** The overall quality of a game (in terms of its design, goals, rules, and mechanics) can be described as playability (Voida & Greenberg, 2012). Playability can also be conceptualized as "the degree to which a game is fun to play and is usable, with an emphasis on the interaction style and plot-quality of the game; the quality of gameplay" (Usability-First, 2009).
- **Ease of use-usability.** Usability is the degree to which a player can understand the basics of a game (e.g., learn its controls) (Pinelle, Wong, & Stach, 2008). In this respect, usability can be viewed as "ease of use", the degree to which a person believes that the use of the given tool effortless (Davis, Bagozzi, & Warshaw, 1989). Ease of use is a factor extensively used in assessing almost all computer applications.
- **Feedback.** Feedback's role is to give players a sense of progress (Cheng, Lin, & She, 2015) and to inform them of the results of their actions (Cheng & Annetta, 2012). This, in turn, allows players to reflect on what they have learned.
- **Narration/storyline.** Even though the narrative introduces the game's fictional background, in the context of SGs, it also provides declarative knowledge for players (Kiili, 2005), keeping them tied up to the game (Couceiro, Papastergiou, Kordaki, & Veloso, 2013).
- **Learning goals.** SGs are goal-directed related to both gaming and learning. Together they can provide engaging and pleasurable experiences on condition they are well-designed, tough, and achievable (Shi & Shih, 2015).
- **Audiovisual adequacy.** Audiovisual features and advanced graphics seize the players' attention and make the game more attractive (Huang, Johnson, & Han, 2013).
- **Realism.** In short, realism refers to how closely real life is replicated within a game. While realism and audiovisual fidelity are closely related, the former has psychological aspects in addition to technical ones (Ravyse et al., 2017). Moreover, how the player interacts with the game also contributes to the game's realism (Mortara, Catalano, Bellotti, Fiucci, Houry-Panchetti, & Petridis, 2014). Thus, realism can be examined as a discrete factor.
- **Relevance to personal interests.** An educational game even if it is enjoyable, it is not necessarily interesting (Squire & Jenkins, 2003). The blending of the content and of the entertainment it provides has to be relevant to the user's own personal interests.
- **Adequacy of the learning material.** The learning content in a serious game has a predominant role. Then again, the learning material has to be seamlessly integrated into the game so as the former not to overshadow the latter and vice versa (Khenissi, Essalmi, & Jemni, 2015). In essence, the learning

content has to be effective without the game being perceived as just an entertaining layer (Mortara et al., 2014). If this is achieved, SGs can engage players even if the learning material is technical and boring, or even when the learning objectives are difficult to achieve.

- **Learning effectiveness.** Learning is the ultimate goal of all SGs and the learning outcomes are their most well-studied factor. While the assessment of the learning outcomes can be based on a variety of methods (e.g., educational objectives' taxonomies, psychomotor, cognitive, and affective domains outlining the learner's capabilities) (Gilbert & Gale, 2007), it is essential to examine whether users perceive SGs as effective (or useful) in terms of the degree they ease knowledge acquisition.

3. THE SCALE'S DEVELOPMENT

The next stage involved the resolution of the measuring issues that arise for each factor. Without a doubt, a number of scales measuring different aspects of users' views do exist, each with its own strengths. Then again, the literature review revealed that for certain SGs' genres, some scales measured a limited number of aspects of the users' views, or contained ambiguous questions. Others did not follow optimal practices for scale development. Having the above deficiencies in mind, the development of the study's scale attempted to follow the most widely accepted steps of scale development and validation, with the first being the item pool generation. More than twenty questionnaires measuring important constructs were reviewed, together with a number of popular questionnaires freely available in the human-computer interaction domain (e.g., the System Usability Scale). The following inclusion or exclusion criteria were applied: (a) to have been tested and validated in studies concerning SGs (or similar applications in case there were no relevant questionnaires), (b) when multiple questions were used for examining a factor, only the questions with high loadings were selected, and (c) when factors were examined using a single question, this question was considered for inclusion only if it loaded exceptionally high on its respective factor.

These sources provided an extensive pool containing more than 400 items. An iterative series of modifications and refinements followed. Redundant and similarly phrased items were removed, poorly worded items were either removed or rephrased, and items that were deemed as not contributing to the assessment of users' views were removed as well. Having a variety of questions and balancing their number (i.e., to avoid the overrepresentation of a factor) was also a consideration. As a result, seventy-two items were retained for the expert review phase in which the content validity was assessed (Worthington & Whittaker, 2006). The pool of questions was translated into Greek by two groups. Each group consisted of one psychologist and one computer science professional with experience in SGs and questionnaire design and development (both proficient in the English language). The resulting versions were then back-translated into English and viewed by a third group of experts. A unified version was obtained through a consensus meeting for assessing the semantic adaptation. Thus, the initial version of the Serious Games Evaluation Scale (SGES) was formulated, having seventy-two items which were supposed to measure thirteen factors. Table 1 presents SGES's factors, the number of items in each factor, and the initial source of these items. All items were presented in a 5-point Likert-type scale, worded "Strongly Agree," "Agree," "Neutral," "Disagree," and "Strongly Disagree."

4. METHOD

As already stated, the purpose of the study was to develop a scale for measuring the users' views/experiences when playing SGs. Having developed the initial scale, the next step was to confirm

Table 1
The items' sources

Factor	Items	Source
Perceived realism	4	Fokides & Atsikpasi, 2018; Witmer & Singer, 1998
Perceived ease of use	6	Brooke, 1996; Phan, Keebler, & Chaparro, 2016; Selwyn, 1997
Perceived playability	6	Brooke, 1996; Phan et al., 2016
Perceived audiovisual adequacy	7	Phan et al., 2016
Perceived narration's adequacy	5	Phan et al., 2016
Perceived feedback's adequacy	4	Fu, Su, & Yu, 2009; Phan et al., 2016
Perceived goal's clarity	4	Fu et al., 2009
Perceived adequacy of the learning material	5	Keller, 1987
Presence/immersion/flow	6	Brockmyer, Fox, Curtiss, McBroom, Burkhart, & Pidruzny, 2009; Fu et al., 2009; IJsselsteijn, De Kort, & Poels, 2013; Novak, Hoffman, & Yung, 2000; Phan et al., 2016; Witmer & Singer, 1998
Enjoyment	6	IJsselsteijn et al., 2013; Keller, 1987; Phan et al., 2016; Tamborini, Bowman, Eden, Grizzard, & Organ, 2010
Motivation	6	Keller, 1987; Martens, Bastiaens, & Kirschner, 2007
Perceived relevance to personal interests	4	Keller, 1987
Perceived learning effectiveness	9	Fokides & Atsikpasi, 2018; Fu et al., 2009; Keller, 1987; Selwyn, 1997

its factorial structure and to validate it. For collecting the necessary data, a project was designed and implemented which lasted from mid-January to mid-March 2018.

4.1. Materials

There are many different SGs genres and their quality is dissimilar as well. On the other hand, the study's objective was not to evaluate a specific SG but the development of a scale able to measure the users' views (either good or bad). In this respect, the game's quality and type were irrelevant. Following this line of thinking, two games developed by Triseum (https://triseum.com/) were selected as the study's material. Although quite different, both can be considered as typical SGs, addressed to young adults (university students). "ARTé Mecenas" is a 2D game, which places users as heads of the Medici family in the tumultuous Italian Renaissance. Its objective is to enable students to appreciate the interconnectedness of local and international economies and how art and art patronage were influenced. Throughout the art history game based on the actual course material, players make decisions, trying to balance their relationship with city-states, merchant factions, and the Catholic Church. The ultimate goal is to build a financial empire and create Renaissance's famous artworks, monuments, and institutions. "Variant: Limits" is a 3D game attempting to enable students to understand, experiment with and appreciate the notion of curriculum-based calculus concepts, such as finite limits, continuity of combined functions, and infinite limits. Players explore a virtual world (a fictitious planet) and manipulate objects within it. Successful understanding of the concepts allows them to open and pass through gates. The objective is to help Equa (the game's main character) to save the planet by solving a series of increasingly complex calculus problems.

4.2. Participants

The sample was students studying at the Department of Audio and Visual Arts, Ionian University, Corfu and at the Department of Primary Education, University of the Aegean, Rhodes, both in Greece.

It has to be stressed that the selection of the above target groups was not a decision taken without consideration. Arté Mecenas' learning subject is history and arts and Variant: Limits' is calculus. In this respect, one option was to select students studying arts' history or advanced maths. Another option was to select participants regardless of their field of studies, in line with the reasoning that led to the selection of the two SGs. Considering both options, it was decided the sample to simulate an audience which is not entirely focused on either of the two SGs but, at the same time, has an interest in playing both. The purpose was to achieve a balanced sample by avoiding over motivated/unmotivated, interested/uninterested participants, without compromising the fact that SGs are addressed, by default, to explicit groups of users. As the curriculum of both groups includes courses related to arts and maths, but these courses are not so specific as was the learning subjects of both games, these groups were considered as ideal for the study's purposes.

A total of 570 students enrolled, in exchange for course credit or for the opportunity to fulfill a course's requirements. They were recruited through a research announcement posted on the Facebook groups these two departments maintain, addressed to anyone interested to participate in the project. There was no initial qualification for research participation.

4.3. Procedure

The participating students were gathered to the Universities' computer labs. They were informed that they were going to play an SG (or two if they were interested in doing so) and complete a questionnaire. They were also informed that the study was conducted on a voluntary basis, that their anonymity was guaranteed, and that completion of the questionnaire was taken as an implicit expression of consent to participation. Their only duty was to play the game(s) for a minimum of two hours (each) and/or complete at least two levels. The introductory/tutorial levels (for familiarizing the players with the interface/controls), did not count as playing the game(s) per se. Immediately after playing the game(s), participants were provided with the questionnaire's link as it was available only online. It has to be noted that although each lab could accommodate around thirty students, it was decided only ten to be present at a time, so as participants to feel more relaxed and have some privacy.

4.4. Data screening-descriptive statistics

The questionnaires were checked for partial and unengaged responses. The number of the valid ones left after this screening was 542. Thus, the final sample consisted of an equal number of students, approximately 23 years old ($M = 22.88$, $SD = .25$), with most coming from the Department of Primary Education ($N = 343$), while 185 were males and 357 were females. Their skills in using computer applications was above the mid-point ($M = 3.58$, $SD = .80$), while their expertise in playing games was average ($M = 3.11$, $SD = 1.10$). In total, ARTé Mecenas was played 303 times and Variant: Limits 239. Scores were obtained by allocating numerical values to responses: "strongly agree" scored 5, "agree" scored 4; "neutral" scored 3; "disagree" scored 2 and "strongly disagree" scored 1. The data were assessed for their normality of distribution. The Shapiro–Wilk tests indicated that they were not normally distributed and they were also negatively skewed (consistent with participants' overall views for the games). On the other hand, the skewness and kurtosis of the data were low (in all cases, skewness $<|1|$ and kurtosis <1), well below the recommended values of skewness $<|2|$ and kurtosis <7 (Finney & DiStefano, 2013). Given that the study was highly exploratory in nature, the data were not transformed, as this allows a better interpretation of the results. Moreover, as factor analysis was to follow, the literature suggested that this analysis can be done even on severely skewed and kurtotic

data (Wang, Fan, & Willson, 1996) and that, in relation to how the Cronbach's alpha is affected, data transformations are not always appropriate when item responses are skewed (Norris & Aroian, 2004).

5. RESULTS ANALYSIS

The initial data set was split into two-random-halves as Exploratory Factor Analysis (EFA), as well as Confirmatory Factor Analysis (CFA) were to be conducted. The EFA was essential for establishing the underlying dimensions between the variables and the latent constructs since the SGES was based on translated and adapted versions of items from multiple sources. For assessing the structure of the seventy-two items in the initial version of SGES, the data were imputed into SPSS 25 and principal axis factor analysis (PAF) with oblique rotation was selected. That is because PAF is more suitable for non-normally distributed data (Costello & Osborne, 2005) and takes into account the covariation between variables (Kline, 2005). Oblique rotation is better suited for research involving human behaviors as it produces more accurate results (Costello & Osborne, 2005). Specifically, the promax rotation (kappa = 4) was used as suggested by others (Matsunaga, 2010).

Item removal was deemed necessary for improving the clarity of the data structure. Several criteria were applied for item(s) deletion: (a) communalities coefficients below .50, (b) factor loadings below |.50|, (c) cross-loadings on two or more factors with loading values greater than |.30|, (d) little contribution to the internal consistency of the scale's scores, (e) low conceptual relevance to a factor, and (f) not conceptually consistent with other items in the same factor (Costello & Osborne, 2005; Tabachnick & Fidell, 2007). Each time an item was removed, the EFA was re-run in order to ensure that there were no major negative effects on the scale's structure. The above process resulted in the removal of nineteen items.

The EFA was then re-run for a final time, with the fifty-three retained items. The data were well suited for factorial analysis because: (a) the sample size ($N = 271$) exceeded Cattell's (1978) rule an absolute minimum of 250 observations, (b) the Kaiser–Meyer–Olkin (KMO) Measure of Sampling Adequacy index was .932, (c) Bartlett's Test of Sphericity was significant ($p < .001$), and (d) the extraction communalities were, in all cases, above the .50 level (the minimum value that was observed was .605) (Hair, Black, Babin, & Anderson, 2010; Tabachnick & Fidell, 2007). Kaiser's (1960) criterion (eigenvalue > 1) suggested that a twelve-factor solution should be considered. For confirming the above suggestion, a parallel analysis was run using O'Connor's method (2000). This analysis confirmed the existence of twelve constructs (Table 2). Moreover, the twelve-factor solution was the most parsimonious and conceptually relevant one. These constructs were named as follows: presence (Pre), enjoyment (Enj), perceived learning effectiveness (PLE), perceived narratives' adequacy (Nar), perceived realism (Real), perceived feedback's adequacy (Feed), perceived audiovisual adequacy (AV), perceived relevance to personal interests (RPI), perceived goals' clarity (Goal), perceived ease of use (PEU), perceived adequacy of the learning material (ALM), and motivation (Mot).

All items loaded high on their respective factors (>.60) while each factor averaged above the .70 recommended level (Hair et al., 2010) (Table 3). Cross-loadings between the retained items was not an issue. Moreover, there were no correlations between the factors greater than .70. The total variance explained by the twelve components was 80.51%. The internal consistency was very good as assessed using Cronbach's alpha (DeVellis, 2016), ranging between .88 and .95 for the constructs, while the overall score was .963 (Table 3, last row).

Following the EFA, the factorial structure was inputted into AMOS 25 for performing CFA using the remaining half of the sample. The internal consistency of the scale and its constructs was reassessed using Cronbach's alpha. It was found that the overall score was $\alpha = .941$ while for the

Table 2
Parallel analysis

Number of factors	Random data eigenvalues	Actual data eigenvalues
1	1.310	18.631
2	1.190	5.772
3	1.100	3.083
4	1.035	2.316
5	.974	2.071
6	.909	1.796
7	.859	1.573
8	.799	1.335
9	.761	1.206
10	.721	1.075
11	.676	.955
12	.634	.875
13	.593	.436
14	.564	.341

Note. The 12th factor is the last one in which the eigenvalues of the actual data exceed the eigenvalues of the random data.

Table 3
Item loadings

Item	Factor loadings											
	AV	PEU	PLE	Enj	ALM	Pre	Real	Mot	Nar	Goal	RPI	Feed
Audiovisual4	**.975**	−.052	−.023	.005	.081	−.021	.002	−.029	−.083	.010	.061	−.079
Audiovisual2	**.900**	−.008	.047	−.046	.085	−.026	−.060	−.063	−.037	−.014	−.025	.084
Audiovisual3	**.893**	.006	.014	.041	.019	−.021	−.017	.010	−.056	−.050	.009	−.040
Audiovisual1	**.891**	−.149	−.059	.096	.014	−.039	−.039	.002	.053	.019	.016	−.004
Audiovisual7	**.757**	.083	.054	−.038	−.082	.056	.013	.030	−.062	.090	−.039	.010
Audiovisual6	**.751**	.082	−.038	.026	−.074	.041	.062	−.040	.067	−.049	−.013	.050
Audiovisual5	**.722**	.059	−.010	−.065	−.071	.013	.125	.090	.153	−.023	−.039	−.022
PEU1	.021	**.883**	.008	−.061	−.123	−.007	−.051	−.013	.106	.042	.068	.021
PEU5	.001	**.835**	−.069	−.072	.068	−.030	−.034	.049	.090	−.003	−.037	.014
PEU6	−.094	**.827**	.033	.110	−.006	−.010	.041	−.159	.060	−.081	.070	.001
PEU3	.039	**.812**	.064	.009	−.013	.048	−.067	−.044	.024	.031	−.014	.014
PEU4	−.037	**.804**	−.129	−.001	.187	−.031	.103	.041	−.129	−.031	−.017	−.022
PEU2	.050	**.783**	.103	.023	.055	−.057	.032	.117	−.107	.011	−.054	−.018
PLE2	−.053	.030	**.933**	−.085	−.014	.053	−.009	−.029	−.128	−.003	.004	.070
PLE1	.002	.060	**.913**	−.009	−.035	.015	.007	.069	−.052	−.045	.040	−.034
PLE4	.048	.086	**.896**	−.006	−.069	.079	−.052	.083	−.075	.022	−.087	.007
PLE3	−.047	−.127	**.777**	.271	−.048	−.086	.008	−.069	.090	−.015	.000	−.035
PLE5	−.004	−.028	**.765**	−.053	.101	−.070	.067	.015	.117	.048	.037	−.030
PLE6	.062	−.079	**.760**	−.026	.110	−.059	.053	−.038	.138	−.044	.025	.010
Enjoyment2	−.042	−.088	−.032	**.965**	.090	−.068	.037	.071	.006	−.069	−.035	.001
Enjoyment3	.079	.032	−.046	**.859**	−.062	.016	−.019	−.025	−.053	.034	.086	.077
Enjoyment1	.010	.112	.073	**.812**	−.066	.094	.000	−.049	.040	.042	−.029	−.066
Enjoyment5	.040	.084	.075	**.763**	.005	.098	−.039	−.045	−.001	.017	−.031	.003
Enjoyment4	−.001	.021	.072	**.753**	.009	.026	−.019	.026	.041	−.032	.099	−.012
Enjoyment6	−.003	−.051	−.012	**.746**	.074	−.081	.016	.121	−.017	.066	−.053	.013

Table 3
(Continued)

Item	Factor loadings											
	AV	PEU	PLE	Enj	ALM	Pre	Real	Mot	Nar	Goal	RPI	Feed
ALM4	−.038	.018	−.003	−.015	**.871**	.003	.052	−.013	.063	.063	.029	−.030
ALM3	−.003	.057	.066	.064	**.836**	−.015	−.006	−.060	.002	−.025	−.072	.054
ALM1	−.047	.059	.056	.036	**.826**	.024	−.072	.046	−.088	.001	.000	−.037
ALM2	.110	.042	−.108	−.004	**.804**	.018	−.012	.040	.031	−.012	.060	−.007
Presence5	−.007	−.033	.065	−.036	.033	**.920**	.065	−.016	.005	−.010	−.022	−.033
Presence1	−.079	−.003	−.086	.106	.085	**.792**	.103	−.013	−.030	−.001	−.029	.004
Presence2	.037	−.025	.093	−.104	−.010	**.774**	−.060	−.015	−.028	.040	.057	.033
Presence3	.030	−.023	−.082	.041	−.072	**.743**	−.051	.079	.069	−.051	−.005	−.018
Realism3	−.015	.043	.038	−.080	−.046	.029	**.927**	.002	.020	.004	.021	−.035
Realism4	.027	−.096	.050	.028	.068	−.020	**.830**	−.021	.052	−.056	−.063	.085
Realism2	.011	.104	−.004	.116	−.095	−.036	**.785**	.003	−.050	.076	−.031	−.037
Presence6	.039	−.040	−.041	−.050	.040	.083	**.743**	−.009	−.030	.019	.093	.011
Motivation4	−.036	−.025	.070	.033	.000	−.007	.015	**.887**	−.041	.009	−.043	.033
Motivation3	.025	.034	.029	−.076	.005	.016	−.041	**.807**	.041	.028	.078	−.053
Motivation1	−.013	−.023	−.073	.167	.010	.034	.002	**.779**	.042	−.042	−.013	.039
Narrative2	−.012	−.013	.011	.029	−.016	−.004	−.026	−.014	**.935**	.016	.016	−.062
Narrative1	.001	−.018	−.032	−.014	.012	.000	.059	.018	**.903**	−.051	−.057	.011
Narrative4	.012	.091	.074	−.049	.006	−.008	−.069	.006	**.807**	.072	−.031	.007
Narrative5	−.007	.038	−.046	.074	.006	.070	.029	.038	**.697**	−.006	.055	.048
Goal1	−.021	.003	−.062	.020	−.040	−.050	.060	.027	−.050	**.933**	.045	−.040
Goal2	−.018	−.011	.000	−.006	−.002	−.028	.039	.034	.025	**.858**	.019	.036
Goal3	.029	−.022	.055	.009	.094	.078	−.082	−.079	.058	**.775**	−.067	.036
RPI4	.000	.034	−.079	−.042	.027	.022	.006	−.023	−.011	.011	**.929**	−.006
RPI1	−.019	.053	.042	.107	−.065	−.056	.001	.051	−.043	.000	**.823**	.001
RPI3	.012	−.097	.158	−.041	.062	.036	.009	.003	.007	.003	**.722**	.020
Feedback3	−.005	.039	.017	.046	−.062	−.012	.038	−.006	−.132	.033	−.055	**.889**
Feedback1	−.030	−.075	−.006	.001	.001	−.029	−.002	.013	.114	.006	.018	**.859**
Feedback2	.037	.087	.002	−.033	.065	.038	−.029	.017	.016	−.040	.072	**.690**
Cronbach's α	.946	.933	.943	.949	.925	.882	.909	.891	.923	.901	.888	.866
Total = .963												

Notes. Extraction method: PAF; rotation method: oblique.

constructs the scores ranged between .85 and .94. Maximum Likelihood (ML) was selected as the estimation method, due to the data not being normally distributed and because other methods (i.e., Asymptotically Distribution Free) require sample sizes of several thousand (Muthén & Kaplan, 1985). Nevertheless, ML is quite robust to mild-to-moderate violations of normality (Matsunaga, 2010), having a negligible negative effect on the quality of the parameter estimates (Brown, 2014). Table 4 and Fig. 1 present the results of this analysis. The standardized estimates ranged from .75 to .93 and were regarded as very good (Hair et al., 2010). All of the R^2 values were above .50, suggesting that the items explained more than half the amount of variance of the latent variable that they belonged.

For model fit assessment, the literature recommended the use of three (and more) fit indices. For the Comparative Fit Index (CFI), values exceeding .95 indicate excellent fit (Hu & Bentler, 1999). For the Root Mean Square Error of Approximation (RMSEA), values less than .06 also indicate a very good model fit (Hu & Bentler, 1999). For the Standardized Root Mean Square Residual (SRMR), values close to zero indicate a perfect fit, while values less than .08 are used as a cut-off point, indicating

Table 4
Results for the measurement model

Item	SE	t	R^2	Item	SE	t	R^2
AV4	.92	–	.84	Real3	.92	–	.85
AV3	.88	31.97	.77	Pre6	.81	25.61	.65
AV1	.87	31.38	.76	Real4	.86	29.23	.75
AV2	.86	30.57	.75	Real2	.80	24.87	.63
AV 7	.77	21.02	.59	Pre1	.79	–	.62
AV6	.81	26.59	.66	Pre5	.93	23.38	.86
AV5	.79	25.14	.62	Pre2	.75	18.59	.56
PLE1	.91	–	.82	Pre3	.76	18.91	.58
PLE2	.83	27.42	.69	Nar4	.86	–	.75
PLE4	.88	31.10	.77	Nar1	.87	27.23	.75
PLE3	.85	29.30	.73	Nar2	.92	30.21	.84
PLE5	.86	30.10	.75	Nar5	.87	27.56	.76
PLE6	.87	30.80	.76	Mot3	.80	–	.63
PEU5	.85	–	.73	Mot4	.89	22.78	.79
PEU4	.77	21.56	.59	Mot1	.87	22.37	.76
PEU3	.86	25.82	.73	Goal2	.90	–	.80
PEU1	.88	27.22	.78	Goal1	.86	26.06	.74
PEU6	.86	25.76	.73	Goal3	.82	24.15	.67
PEU2	.83	24.48	.69	RPI4	.83	–	.69
Enj2	.87	–	.76	RPI3	.84	22.23	.70
Enj6	.78	26.05	.61	RPI1	.88	23.44	.78
Enj5	.87	28.03	.75	Feed3	.84	–	.71
Enj3	.91	31.31	.83	Feed2	.83	22.19	.69
Enj4	.91	30.74	.82	Feed1	.86	23.00	.74
Enj1	.91	30.79	.82				
ALM4	.91	–	.84				
ALM1	.84	27.06	.70				
ALM3	.85	27.74	.72				
ALM2	.82	25.97	.67				

Notes. –: This value was fixed at 1.00 for model identification purposes; SE: standardized estimate.

excellent fit (Hu & Bentler, 1999; McDonald & Ho, 2002). Finally, experts recommended not to rely on the chi-square test statistic when the sample size exceeds 200 cases, as it has the tendency to indicate statistically significant differences (Hu & Bentler, 1999). Instead, they recommended the use of the minimum discrepancy divided by its degrees of freedom (χ^2/df), with acceptable values ranging between 1 and 3. The results revealed that the hypothesized twelve-factor model had an excellent fit when the above indices were examined, as presented in Table 5 (second column).

The hypothesized twelve-factor model was compared against three alternative models in terms of overall model fit. The eleven-factor model merged enjoyment and feedback, while the ten-factor model merged enjoyment, narrative, and feedback. A one-factor model was also used as a baseline. All models had the same number of cases and observed variables or items. On the basis of the results, as presented in Table 5, it is evident that the twelve-factor solution had the best overall fit indices.

SGES's convergent validity was checked by measuring the average variance extracted (AVE) (Table 6). The AVE in all but one case was above the .70 level as suggested by Hu and Bentler (1999). Although the AVE of Pre fell slightly below the .70 threshold, given that all the other indices were more than satisfactory, it was considered an acceptable deviation from the recommended values. Moreover,

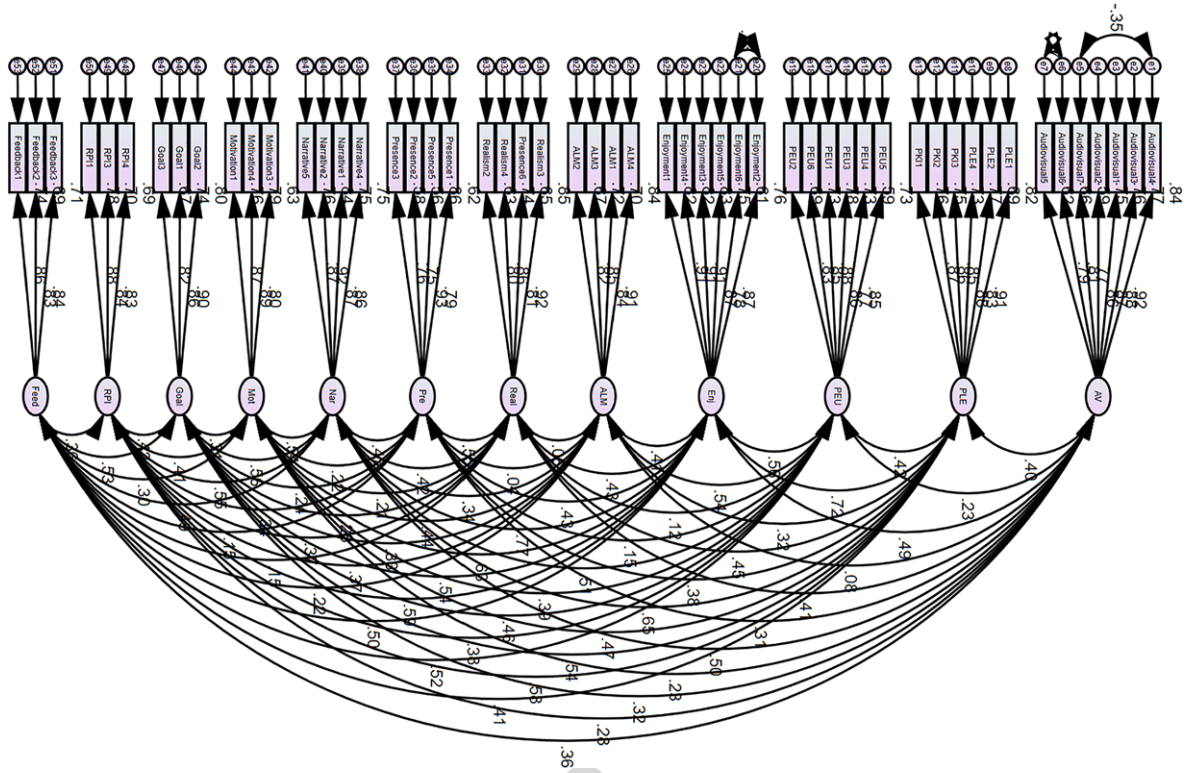

Fig. 1. Results of the CFA.

Table 5
Model fit measures' comparisons

Measure	12 factor model estimates	11 factor model estimates	10 factor model estimates	1 factor model estimates	Threshold
χ^2/df	1.72	2.41	3.49	11.36	Between 1 and 3
CFI	.964	.929	.874	.456	>.95
SRMR	.042	.053	.073	.131	<.08
RMSEA	.036	.051	.068	.138	<.06

reliability was also evident, as all critical ratios were above .70 (Hancock, 2001). The presence of discriminant validity was evaluated by comparing the square root of the AVE for any given factor with the correlations between this factor and all other factors (Hu & Bentler, 1999). It was found that the variance shared between a factor and any other factor was less than the variance that the construct shared with its measures. Thus, discriminant validity was satisfactory in all cases.

Finally, both the EFA and the CFA were re-run, but this time the data were split according to the SG that was used. The purpose was to test whether the data coming from each SG separately had an impact on the scale's factorial structure, validity, and reliability. The results indicated that the factorial structure remained unchanged in both SGs. In addition, the model fit indices were excellent for both games (SG1: χ^2/df = 1.86, CFI = .95, SRMR = .044, RMSEA = .037; SG2: χ^2/df = 1.71, CFI = .96, SRMR = .046, RMSEA = .038). No significant changes whatsoever were observed on the scale's results regarding its convergent and discriminant validity.

Table 6
Convergent and discriminant validity

	CR	AVE	AV	PLE	PEU	Enj	ALM	Real	Pre	Nar	Mot	Goal	RPI	Feed
AV	.946	.714	**.845**											
PLE	.948	.752	.401	**.867**										
PEU	.935	.707	.228	.420	**.841**									
Enj	.951	.765	.488	.724	.560	**.875**								
ALM	.916	.731	.080	.322	.543	.413	**.855**							
Real	.911	.721	.408	.449	.120	.426	.065	**.849**						
Pre	.883	.655	.310	.384	.150	.427	.041	.554	**.809**					
Nar	.932	.775	.504	.648	.506	.772	.336	.424	.470	**.881**				
Mot	.890	.729	.228	.469	.389	.627	.444	.214	.217	.541	**.854**			
Goal	.895	.739	.323	.544	.460	.538	.333	.281	.244	.560	.313	**.860**		
RPI	.886	.721	.284	.576	.383	.586	.372	.366	.343	.552	.412	.420	**.849**	
Feed	.882	.713	.365	.405	.523	.501	.218	.147	.150	.495	.302	.527	.358	**.844**

Notes. CR: Critical ratio; AVE: Average Variance Extracted; diagonal in bold: square root of AVE extracted from observed variables; off-diagonal: correlations between constructs.

In conclusion, the results of the EFA and CFA established the instrument's factorial structure and demonstrated that it had satisfactory validity and reliability. The final version of the SGES with the fifty-three retained items and their respective constructs is presented in the Appendix.

6. DISCUSSION

The literature proposes several assessment methods for serious games. According to Ifenthaler, Eseryel, and Ge (2012) there are three distinct types of evaluation: (a) game scoring (e.g., targets achieved, obstacles or time needed to complete a task or an iteration), (b) external assessment (e.g., interviews, tests or surveys), and (c) embedded or internal assessment (e.g., learner's behavior such as clickstreams or log files). Faizan, Löffler, Heininger, Utesch, and Krcmar (2019) in their literature review classified the evaluation methods in accordance to three phases: (a) pre-game, (b) in-game, and (c) post-game, in order to assess their learning effectiveness. In general, researchers collect data through questionnaires regarding respondents' opinions, attitudes, feelings, and perceptions on a particular matter. Questionnaires are the most frequently used evaluation tool during the pre- and post-game phases. However, the in-game assessment is also important because this type of evaluation provides instant feedback about the learning process. Such investigations traditionally include counting the number of mistakes, checkpoints, in-game performance tracking, storyboarding, game experience questionnaire while playing, monitoring students' progress, think-aloud protocols, self-reports, interviews, and unobtrusive observation. Along with observation, useful tools also include video recordings, detection of facial electromyography activity, measuring the effect of emotion in multiple vocal cues and the physiological activity measures (e.g., heart rate, blood volume pressure, and skin conductance). Currently, a diversity of wearable sensors is used for such measurements (Rebelo, Noriega, Duarte, & Soares, 2012). The usage of these tools demands subjects' agreement taking into account all the ethical commitments of the researchers. Others advance further the investigation domain by combining interaction parameters, facial expression recognition, and/or audio analysis in combination with machine learning techniques (Bartlett, Hager, Ekman, & Sejnowski, 1999; Pham, Kim, Lu, Jung, & Won, 2019; Wu & Lin, 2018; Zeng, Pantic, Roisman, & Huang, 2009). The main advantage of these nonverbal methods is that measurements are more objective compared with sub-

jective data derived by questionnaires and self-reports. Post-game phase evaluation includes questionnaires, measure learner's knowledge after the game. Although all the above methods are valid (and valuable), each has its own field of use depending on the researchers' needs. Consequently, in our case, a post-game questionnaire was chosen as an assessment method, because the objective of our work was to formulate a scale to holistically assess serious games.

Indeed, both the industry and researchers are in need of a comprehensive scale which is psychometrically validated and suitable for evaluation purposes. Toward this end, a new scale called SGES was developed based on a rigorous multistage system of scale development and validation. In this pursuit, a sufficient number of resources (e.g., existing scales and general-purpose scales) concerning SGs' evaluation were screened and generated the initial item pool. The item pool underwent a series of iterative phases of modifications and expert reviews before being pilot-tested. Once refined, the scale was administered to a large sample of 542 university students, who evaluated two SGs. For data analyses, EFA and CFA were performed in order to uncover the underlying factors and for validating the scale.

Out of the initial seventy-two items and the inclusion of thirteen factors, nineteen items and one factor (perceived playability) had to be dropped during the EFA. The retained factors were: presence, enjoyment, perceived learning effectiveness, perceived narratives' adequacy, perceived realism, perceived feedback's adequacy, perceived audiovisual adequacy, perceived relevance to personal interests, perceived goals' clarity, perceived ease of use, perceived adequacy of the learning material, and motivation. A number of reasons might have contributed to this outcome. All questions were originally in English; some items might have been poorly translated into Greek or participants might have found their meaning ambiguous. Indeed, in the final scale, there is an item which was supposed to measure presence (Pre6: "There were times when the virtual objects seemed to be as real as the real ones"), but it proved to be the question with the third strongest loading on perceived realism. The most significant reason for dropping items was the rules that were applied for item and factor retention. In general, Hair et al.'s (2010) recommendation for high items' loading ($>.60$) and high factors' average ($>.7$) was followed. Perceived playability's items proved to be the most problematic ones; they either loaded low on perceived ease of use or there were significant cross-loadings with the above factor as well as with other factors (e.g., perceived feedback's adequacy). On the basis of the results, it seems that participants viewed ease of use, usability, and playability as a single concept and merged them in one factor under the name "perceived ease of use".

Even so, the final scale does not contain factors represented with less than three items. If this was the case, these factors might have been considered as unstable (Costello & Osborne, 2005; Raubenheimer, 2004). Actually, only four factors are measured using three items (motivation, perceived relevance to personal interests, perceived goals' clarity, and perceived feedback's adequacy), while all the other factors are measured at least with four. Yet, the total variance explained by the fifty-three items was 80.51%, which is more than satisfactory (Hair et al., 2010). Moreover, SGES's reliability and internal consistency, as a whole and per construct, was well above the .70 threshold ($\alpha = .88$ to .96) which is considered very good for social science research (DeVellis, 2016). Therefore, it can be argued that SGES is a quite balanced scale since no factor is overrepresented.

The same holds true for SGES's convergent and discriminant validity as no problems were noted during the CFA. Indeed, results obtained from the CFA demonstrated that the model fit indices were exceptionally good and that the same applied for its discriminant validity. One minor deviation from the recommended threshold for convergent validity was noted in one factor, namely presence (Pre $=$ .655, recommended value $>$ .70), but it was considered acceptable as it does not affect the scale's overall convergent validity.

Taking together the above, it can be concluded that SGES seems to be a quite robust scale and short, in terms of how many items it has and how many factors it measures. The estimated time needed for completing it is between ten to fifteen minutes.

6.1. Implications for research and practice

Past research provided evidence that SGs can be effective learning tools (Connolly et al., 2012; de Freitas, 2018; Erhel & Jamet, 2019; Lamb et al., 2018) and that they can be useful in a wide range of learning subjects and educational scenarios (Girard et al., 2013). While this holds true, until now, there was no common consensus on how to measure the users' views for these applications. Though studies scrutinizing the impact of certain factors on the learning outcomes do exist, each analyzed either one (e.g., Khan & Webster, 2017) or a different set of factors (e.g., All et al., 2015; Marsh, 2011; Steiner et al., 2015) and each used different instruments for validating these factors. Very few studies implied that a larger number of factors should be considered when evaluating SGs but did not suggest (or validate) an instrument for examining them (e.g., Calderón & Ruiz, 2015). On the other hand, in this study, by reviewing the relevant literature, twelve subjective factors were located and a scale for measuring them was developed and tested. The scale's substantial internal consistency, stability, and validity, are indicators that, indeed, these factors shape the user's views when playing SGs. Thus, the study's contribution to research is that it proposes an instrument for measuring multiple factors. This, in turn, can lead to the establishment of a much-desired common methodology for measuring SGs effectiveness (Serrano-Laguna et al., 2018). In addition, as all the factors included in SGES are subjective ones, it can provide a better understanding of what shapes users' experiences when playing SGs. As presented in a preceding section, while the term "user experience" is ill-defined (Buck et al., 2018; Koeffel et al., 2010), it is acknowledged that it is an important aspect of human-computer interaction (Lallemand et al., 2015) and plays an important role in the success of educational applications.

Additionally, there are two reasons that render the scale a flexible tool. First, it can be used for assessing a variety of SGs. That is because two quite different SGs were used and the analysis provided evidence that, in both cases, the attributes of the scale remained unchanged. In this respect, SGES can be considered as a first step in overcoming a major problem highlighted by others, namely, the generalizability of the evaluation methods (e.g., Ravyse et al., 2017). Second, it can be assumed, within reasonable limits, that the scale has a modular structure, meaning that factors can be excluded without altering the instrument's validity. The strong items' loadings on their respective factors and their minimal cross-correlations justify this assumption. As a result, the scale can be used in a variety of situations, depending on the researchers' needs, allowing them to assess different affordances, individually or together. For example, if an SG under evaluation has no narrative components, researchers (or practitioners for that matter) can remove the perceived narration's adequacy component from the scale.

The practical applications of the SGES derive from its robust structure and satisfactory attributes. It is suitable for a wide variety of target groups and SGs, as long as the games examined possess the aspects that are being measured and players are capable to use them. Indeed, two different SGs were used and the sample in which it was administered consisted of university students with mixed computer and game-playing competences (ranging from novice to expert level, see section "Data screening-descriptive statistics"). Thus, experts from the game industry can average (or sum) the scores in each factor and/or obtain a composite score indicating the overall users' views and experiences. These scores can then be used for comparing different SGs (either of the same or different genre). Alternatively, they can do the same for comparing different versions of the same SG and determine if the

latest version is perceived to be an improvement compared to the previous ones. Educators can also benefit from the use of SGES in a similar fashion. By administering it to their students, they can determine which aspects of an SG contributed either positively or negatively to the learning outcomes. Also, within specific and focused application fields, by administering different systems to students, educators can determine which aspects of an SG contributed either positively or negatively to the learning outcomes, enabling them to combine features and create superior systems before initiating their own developmental process.

Although developers and educators can independently benefit from the use of SGES, our view is that this process has to be implemented in a combined fashion. A single run of the evaluation process reveals to both developers and educators the strong and weak parts of the SG. Thus, this decomposition that is offered inherently by the method can help speed-up the iterative or recursive lifecycle of the environment as both parties can identify and improve the system (interaction design, usability, aesthetics) and content deficiencies altogether.

7. CONCLUSION

The study has several limitations that need to be addressed. The trustworthiness of participants' responses in questionnaires is always a concern. University students coming from two quite different disciplines of study took part in the study; the sample might not be representative of the intended audience of the SGs that were used. Participants were asked to play the games for two hours. This length of time might not be sufficient for users to form a comprehensive view of the SGs. Moreover, the study was based on the assumption that SGs' quality and genre are not important as it sought to develop a scale for examining how users view them. On the other hand, it is unknown how participants might have reacted if different SGs were used. Since the study was conducted in a controlled environment with other participants present, it is possible their views to have been affected; if the SGs were played in a more relaxed environment (e.g., at home) the results might have been different. Another limitation is the scale's newness. As it was just developed, there is no information on SGES's scoring standard.

The above limitations may serve as directions for future research. Since SGES is a new scale, further validations are definitely needed that will provide evidence for its validity and reliability. Different target groups (e.g., in terms of age, studies, and level of education), as well as a larger variety of SGs, will demonstrate if and how its structure is affected. Additional factors that shape the learning and the gaming experience can be considered. A research path the authors are already planning to explore is to compare the subjective factors the scale measures with data that measure actual learning (e.g., through knowledge acquisition tests). This will provide evidence on how (and to what extent) the factors interact with each other and what impact they have on the learning outcomes.

Finally, an important factor that needs to be further examined are the different playing patterns/intentions and the resulting game adaptations that have to occur in order to render SGs useful for each user. For example, users may play a game without prior knowledge on the subject matter, while others may play it after having studied the theory and related texts. Under certain circumstances, gamers may use hints and tips to progress faster within the game (a common practice across players), while others may not use this feature. In future revisions of SGES, questions related to the above can be considered for inclusion. Moreover, it would be interesting to examine whether it is possible to integrate the scale into an SG allowing the latter to dynamically adjust, leading to a personalized and improved gamers' experience.

In sum, despite the above limitations, the study contributes to the relevant literature by providing a tool that simultaneously assesses many factors that shape the users' views when playing SGs. Thus, the study's results might prove useful to educators, researchers, and developers in planning their lessons, in understanding the interactions between these factors, and for designing even more effective SGs.

APPENDIX

Table A.1
The final version of SGES

Factor	Item
Presence	I was deeply concentrated in the game
	If someone was talking to me, I couldn't hear him
	I forgot about time passing while using the game
	I felt detached from the outside world while using the game
Enjoyment	I think the game was fun
	I felt bored while using the game*
	I enjoyed using the game
	I really enjoyed studying with this game
	It felt good to successfully complete the tasks in this game
	I felt frustrated*
Perceived learning effectiveness	I felt that this game can ease the way I learn
	This game was a much easier way to learn compared to the usual teaching
	This game made learning more interesting
	I felt that the game increased my knowledge
	I felt that I caught the basics of what I was taught with this game
	I will definitely try to apply the knowledge I learned with this game
Perceived narratives'	I was captivated by the game's story from the beginning
	I enjoyed the fantasy or story provided by the game
	I could clearly understand the game's story
	I was very interested in seeing how the events in the game will unfold
Perceived realism	When interacting with the virtual objects, these interactions seemed like real ones
	There were times when the virtual objects seemed to be as real as the real ones
	The virtual objects seemed like real objects to me
	When I used the game, the virtual world was more real than the real world
Perceived feedback's	I received immediate feedback on my actions
	I was notified of new tasks immediately
	I received information on my success (or failure) on the intermediate goals immediately
Perceived audiovisual adequacy	I enjoyed the sound effects in the game
	I think the game's audio fits the mood or style of the game
	I felt the game's audio (e.g., sound effects, music) enhanced my (gaming) experience
	I enjoyed the music in the game
	I enjoyed the game's graphics
	I think the game was visually appealing
	I think the game's graphics fit the mood or style of the game
Perceived relevance to	The content of this game was relevant to my interests
	I could relate the content of this game to things I have seen, done, or thought about in my own life
	It is clear to me how the content of the game is related to things I already know

Table A.1
(Continued)

Factor	Item
Perceived goals' clarity	The game's goals were presented at the beginning of the game
	The game's goals were presented clearly
	The intermediate goals were presented at the beginning of each scene
Perceived ease of use	I think it was easy to learn how to use the game
	I found the game unnecessarily complex*
	I imagine that most people will learn to use this game very quickly
	I needed to learn a lot of things before I could get going with this game*
	I felt that I needed help from someone else in order to use the game because It was not easy for me to understand how to control the game*
	It was easy for me to become skillful at using the game
Perceived adequacy of the learning material	In some cases, there was so much information that it was hard to remember the important points*
	The exercises in this game were too difficult*
	I could not really understand quite a bit of the material in this game*
	The good organization of the content helped me to be confident that I would learn this material
Motivation	This game did not hold my attention*
	When using the game, I did not have the impulse to learn more about the learning subject*
	The game did not motivate me to learn*

Note. * = Item for which its scoring was reversed.

REFERENCES

Abt, C.C. (1970). *Serious Games*. New York: Viking Press.

All, A., Castellar, E.P.N. & Van Looy, J. (2015). Towards a conceptual framework for assessing the effectiveness of digital game-based learning. *Computers & Education*, *88*, 29–37. doi:10.1016/j.compedu.2015.04.012.

Bartlett, M.S., Hager, J.C., Ekman, P. & Sejnowski, T.J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, *36*(2), 253–263. doi:10.1017/S0048577299971664.

Blumberg, F.C., Almonte, D.E., Anthony, J.S. & Hashimoto, N. (2013). Serious games: What are they? What do they do? Why should we play them? In K.E. Dill (Ed.), *The Oxford Handbook of Media Psychology* (pp. 334–351). Oxford University Press.

Brockmyer, J.H., Fox, C.M., Curtiss, K.A., McBroom, E., Burkhart, K.M. & Pidruzny, J.N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, *45*(4), 624–634. doi:10.1016/j.jesp.2009.02.016.

Brooke, J. (1996). SUS – a quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7.

Brown, T.A. (2014). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). New York, NY: Guilford Press.

Buck, R., Khan, M., Fagan, M. & Coman, E. (2018). The User Affective Experience Scale: A measure of emotions anticipated in response to pop-up computer warnings. *International Journal of Human–Computer Interaction*, *34*(1), 25–34. doi:10.1080/10447318.2017.1314612.

Calderón, A. & Ruiz, M. (2015). A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education*, *87*, 396–422. doi:10.1016/j.compedu.2015.07.011.

Cattell, R.B. (1978). *The Scientific Use of Factor Analysis*. New York: Plenum. doi:10.1007/978-1-4684-2262-7.

Cheng, M.T. & Annetta, L. (2012). Students' learning outcomes and learning experiences through playing a Serious Educational Game. *Journal of Biological Education*, *46*(4), 203–213. doi:10.1080/00219266.2012.688848.

Cheng, M.T., Lin, Y.W. & She, H.C. (2015). Learning through playing Virtual Age: Exploring the interactions among student concept learning, gaming performance, in-game behaviors, and the use of in-game characters. *Computers & Education*, *86*, 18–29. doi:10.1016/j.compedu.2015.03.007.

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. doi:10.1016/j.compedu.2012.03.004.

Costello, A.B. & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9.

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. doi:10.1007/s10639-011-9179-3.

Davis, F.D., Bagozzi, R.P. & Warshaw, P.R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982–1003. doi:10.1287/mnsc.35.8.982.

de Freitas, S. (2018). Are games effective learning tools? A review of educational games. *Journal of Educational Technology & Society*, *21*(2), 74–84.

DeVellis, R.F. (2016). *Scale Development: Theory and Applications* (Vol. 26). Sage Publications.

Erhel, S. & Jamet, E. (2019). Improving instructions in educational computer games: Exploring the relations between goal specificity, flow experience and learning outcomes. *Computers in Human Behavior*, *91*, 106–114. doi:10.1016/j.chb.2018.09.020.

Ermi, L. & Mäyrä, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. *Worlds in Play: International Perspectives on Digital Games Research*, *37*(2), 37–53.

Faizan, N., Löffler, A., Heininger, R., Utesch, M. & Krcmar, H. (2019). Classification of evaluation methods for the effective assessment of simulation games: Results from a literature review. *International Journal of Engineering Pedagogy*, *9*(1), 19–33. doi:10.3991/ijep.v9i1.9948.

Finney, S.J. & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G.R. Hancock and R.O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (2nd ed., pp. 269–314). Charlotte, NC: IAP.

Fokides, E. & Atsikpasi, P. (2018). Factors affecting primary school students' learning outcomes when using MUVEs. Development and validation of a scale. In J.Y. Qian (Ed.), *Integrating Multi-User Virtual Environments in Modern Classrooms* (pp. 185–206). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-3719-9.ch009.

Fokides, E., Atsikpasi, P., Kaimara, P. & Deliyannis, I. (2019). Serious games: Which factors players consider important for their learning and gaming experience? Manuscript submitted for publication.

Fu, F.L., Su, R.C. & Yu, S.C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education*, *52*(1), 101–112. doi:10.1016/j.compedu.2008.07.004.

Garris, R., Ahlers, R. & Driskell, J.E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, *33*(4), 441–467. doi:10.1177/1046878102238607.

Gilbert, L. & Gale, V. (2007). *Principles of E-Learning Systems Engineering*. Elsevier. doi:10.1533/9781780631196.

Girard, C., Ecalle, J. & Magnan, A. (2013). Serious games as new educational tools: How effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning*, *29*(3), 207–219. doi:10.1111/j.1365-2729.2012.00489.x.

Hair, J.F., Black, W.C., Babin, B.J. & Anderson, R.E. (2010). *Multivariate Data Analysis: International Version* (7th ed.). New Jersey: Pearson.

Hancock, G.R. (2001). Rethinking construct reliability within latent variable systems. In *Structural Equation Modeling: Present and Future* (pp. 195–216). Scientific Software International.

Hersh, M. & Leporini, B. (2018). Serious games, education and inclusion for disabled people. *British Journal of Educational Technology*, *49*(4), 587–595. doi:10.1111/bjet.12650.

Hu, L.T. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118.

Huang, W.D., Johnson, T.E. & Han, S.H.C. (2013). Impact of online instructional game features on college students' perceived motivational support and cognitive investment: A structural equation modeling study. *The Internet and Higher Education*, *17*, 58–68. doi:10.1016/j.iheduc.2012.11.004.

Huang, W.H., Huang, W.Y. & Tschopp, J. (2010). Sustaining iterative game playing processes in DGBL: The relationship between motivational processing and outcome processing. *Computers & Education*, *55*(2), 789–797. doi:10.1016/j.compedu.2010.03.011.

Ifenthaler, D., Eseryel, D. & Ge, X. (2012). Assessment for game-based learning. In *Assessment in Game-Based Learning* (pp. 1–8). New York, NY: Springer. doi:10.1007/978-1-4614-3546-4_1.

IJsselsteijn, W., De Kort, Y.A.W. & Poels, K. (2013). *The Game Experience Questionnaire*. Eindhoven: Technische Universiteit Eindhoven. Retrieved from https://pure.tue.nl/ws/files/21666907/GaMediume_Experience_Questionnaire_English.pdf.

Ivory, J.D. & Kalyanaraman, S. (2007). The effects of technological advancement and violent content in video games on players' feelings of presence, involvement, physiological arousal, and aggression. *Journal of Communication*, *57*(3), 532–555. doi:10.1111/j.1460-2466.2007.00356.x.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151. doi:10.1177/001316446002000116.

Keller, J.M. (1987). *IMMS: Instructional Materials Motivation Survey*. Florida State University.

Khan, A. & Webster, J. (2017). Digital game narrative quality: Developing a measure. In *Proceedings of the Thirty-Eighth International Conference on Information Systems*, Seoul.

Khenissi, M.A., Essalmi, F. & Jemni, M. (2015). Comparison between serious games and learning version of existing games. *Procedia – Social and Behavioral Sciences*, *191*, 487–494. doi:10.1016/j.sbspro.2015.04.380.

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *The Internet and Higher Education*, *8*(1), 13–24. doi:10.1016/j.iheduc.2004.12.001.

Kline, R.B. (2005). *Principles and Practice of Structural Equation Modeling* (2nd ed.). New York: Guilford Press.

Koeffel, C., Hochleitner, W., Leitner, J., Haller, M., Geven, A. & Tscheligi, M. (2010). Using heuristics to evaluate the overall user experience of video games and advanced interaction games. In R. Bernhaupt (Ed.), *Evaluating User Experience in Games* (pp. 233–256). London: Springer. doi:10.1007/978-1-84882-963-3_13.

Lallemand, C., Gronier, G. & Koenig, V. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in Human Behavior*, *43*, 35–48. doi:10.1016/j.chb.2014.10.048.

Lamb, R.L., Annetta, L., Firestone, J. & Etopio, E. (2018). A meta-analysis with examination of moderators of student cognition, affect, and learning outcomes while using serious educational games, serious games, and simulations. *Computers in Human Behavior*, *80*, 158–167. doi:10.1016/j.chb.2017.10.040.

Marsh, T. (2011). Serious games continuum: Between games for purpose and experiential environments for purpose. *Entertainment Computing*, *2*(2), 61–68. doi:10.1016/j.entcom.2010.12.004.

Martens, R., Bastiaens, T. & Kirschner, P.A. (2007). New learning design in distance education: The impact on student perception and motivation. *Distance Education*, *28*(1), 81–93. doi:10.1080/01587910701305327.

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research*, *3*(1), 97–110. doi:10.21500/20112084.854.

McDonald, R.P. & Ho, M.R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82. doi:10.1037/1082-989X.7.1.64.

Mortara, M., Catalano, C.E., Bellotti, F., Fiucci, G., Houry-Panchetti, M. & Petridis, P. (2014). Learning cultural heritage by serious games. *Journal of Cultural Heritage*, *15*(3), 318–325. doi:10.1016/j.culher.2013.04.004.

Muratet, M., Viallet, F., Torguet, P. & Jessel, J. (2009). Une ingénierie pour jeux sérieux [Engineering for serious games]. In *Proceedings of Conférence EIAH-Atelier Jeux Sérieux* (pp. 53–63). France: Le Mans.

Muthén, B. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171–189. doi:10.1111/j.2044-8317.1985.tb00832.x.

Norris, A.E. & Aroian, K.J. (2004). To transform or not transform skewed data for psychometric analysis: That is the question! *Nursing Research*, *53*(1), 67–71. doi:10.1097/00006199-200401000-00011.

Novak, T.P., Hoffman, D.L. & Yung, Y.-F. (2000). Measuring the customer experience in online environments: A structural modeling approach. *Marketing Science*, *19*(1), 22–42. doi:10.1287/mksc.19.1.22.15184.

O'Brien, S. (2016). *Why User Experience Design Is Critical to Driving and Maintaining User Engagement and Motivation for Online and Mobile Educational Tools?* Proceedings of the Future of Education Conference (Vol. 328). Libreriauniversitaria Edizioni.

O'Connor, B.P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, *32*, 396–402. doi:10.3758/BF03200807.

Pham, T.T.D., Kim, S., Lu, Y., Jung, S.W. & Won, C.S. (2019). Facial action units-based image retrieval for facial expression recognition. *IEEE Access*, *7*, 5200–5207. doi:10.1109/ACCESS.2018.2889852.

Phan, M.H., Keebler, J.R. & Chaparro, B.S. (2016). The development and validation of the Game User Experience Satisfaction Scale (GUESS). *Human Factors*, *58*(8), 1217–1247. doi:10.1177/0018720816669646.

Pinelle, D., Wong, N. & Stach, T. (2008). Heuristic evaluation for games: Usability principles for video game design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1453–1462). New York: ACM. doi:10.1145/1357054.1357282.

Raubenheimer, J. (2004). An item selection procedure to maximise scale reliability and validity. *SA Journal of Industrial Psychology*, *30*(4), 59–64. doi:10.4102/sajip.v30i4.168.

Ravyse, W.S., Blignaut, A.S., Leendertz, V. & Woolner, A. (2017). Success factors for serious games to enhance learning: A systematic review. *Virtual Reality*, *21*(1), 31–58. doi:10.1007/s10055-016-0298-4.

Rebelo, F., Noriega, P., Duarte, E. & Soares, M. (2012). Using virtual reality to assess user experience. *Human Factors*, *54*(6), 964–982. doi:10.1177/0018720812465006.

Selwyn, N. (1997). Students' attitudes toward computers: Validation of a computer attitude scale for 16–19 education. *Computers & Education*, *28*, 35–41. doi:10.1016/S0360-1315(96)00035-8.

Serrano-Laguna, Á., Manero, B., Freire, M. & Fernández-Manjón, B. (2018). A methodology for assessing the effectiveness of serious games and for inferring player learning outcomes. *Multimedia Tools and Applications*, *77*(2), 2849–2871. doi:10.1007/s11042-017-4467-6.

Shi, Y.R. & Shih, J.L. (2015). Game factors and game-based learning design model. *International Journal of Computer Games Technology*, *2015*, 11 pp. doi:10.1155/2015/549684.

Squire, K. & Jenkins, H. (2003). Harnessing the power of games in education. *Insight*, *3*(1), 5–33.

Steiner, C., Hollins, P., Kluijfhout, E., Dascalu, M., Nussbaumer, A., Albert, D. & Westera, W. (2015). Evaluation of serious games: A holistic approach. Retrieved from http://dspace.ou.nl/bitstream/1820/6139/1/RAGE_ICERI2015_final.pdf.

Tabachnick, B.G. & Fidell, L.S. (2007). *Using Multivariate Statistics*. Boston: Pearson Education Inc.

Tamborini, R., Bowman, N.D., Eden, A., Grizzard, M. & Organ, A. (2010). Defining media enjoyment as the satisfaction of intrinsic needs. *Journal of Communication*, *60*(4), 758–777. doi:10.1111/j.1460-2466.2010.01513.x.

Usability-First (2009). Playability definition. Retrieved from http://www.usabilityfirst.com/glossary/playability/.

Voida, A. & Greenberg, S. (2012). Console gaming across generations: Exploring intergenerational interactions in collocated console gaming. *Universal Access in the Information Society*, *11*(1), 45–56. doi:10.1007/s10209-011-0232-1.

Wang, L., Fan, X. & Willson, V.L. (1996). Effects of nonnormal data on parameter estimates and fit indices for a model with latent and manifest variables: An empirical study. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(3), 228–247. doi:10.1080/10705519609540042.

Witmer, B.G. & Singer, M.J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence*, *7*(3), 225–240. doi:10.1162/105474698565686.

Worthington, R.L. & Whittaker, T.A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838. doi:10.1177/0011000006288127.

Wu, B.F. & Lin, C.H. (2018). Adaptive feature mapping for customizing deep learning based facial expression recognition model. *IEEE Access*, *6*, 12451–12461. doi:10.1109/ACCESS.2018.2805861.

Zeng, Z., Pantic, M., Roisman, G.I. & Huang, T.S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58. doi:10.1109/TPAMI.2008.52.