Journal on Interactive Systems, 2021, 12:1, doi: 10.5753/jis.2021.1907

Planning the Design and Execution of Student Performance Assessment in Serious Games

Rháleff N. R. Oliveira [Universidade Federal do ABC | *rhaleff.nascimento@ufabc.edu.br* **]**
Rafaela V. Rocha [Universidade Federal do ABC | *rafaela.rocha@ufabc.edu.br* **]**
Denise H. Goya [Universidade Federal do ABC | *denise.goya@ufabc.edu.br* **]**

Abstract

Serious Games (SGs) are used to support knowledge acquisition and skill development. For this, there is a need to measure the results achieved (both during and after students play) to ensure the game effectiveness. In this context, the aim is to develop and evaluate the AvaliaJS, a conceptual model to structure, guide and support the planning of the design and execution of the student's performance assessment in SGs. AvaliaJS has two artifacts: a canvas model, for high-level planning, and an assessment project document, for more detailed specifications of the canvas. To analyze and exemplify the use of the model, the artifacts were applied to three ready-made games as a proof of concept. In addition, the quality of AvaliaJS was evaluated by experts in SGs development and assessment using a questionnaire. The results of experts' answers confirm a good internal consistency (*Cronbach's alpha* α = 0.87) which indicates that AvaliaJS is correct, authentic, consistent, clear, unambiguous and flexible. However, the model will need to be validated during the process of creating a new game to ensure its usability and efficiency. In general, AvaliaJS can be used to support the team in the planning, documentation and development of artifacts and data collection in SGs, as well as in the execution of the assessment, learning measurement and constant and personalized feedback for students.

Keywords: *Serious Games, Conceptual Model, Human Performance Assessment, Evaluation*

1 Introduction

There are two main points in Serious Games (SGs): (1) combination of games with one or more functions, such as broadcasting a message, providing training or promoting data exchange; and (2) application in a specific domain, such as defense, training, education, healthcare, and not just in the entertainment area (Zyda, 2005; Alvarez; Djaouti, 2012). Because they have educational functions, serious games are characterized by the focus on teaching a certain content, helping to explore new skills, disseminating concepts and seeking to change an attitude (Dempsey *et al.*, 1996).

The effectiveness of SGs refers to the ability to ensure that learning happened, by offering an effective assessment of student performance (in assisting learning and skills development) (Salas *et al.*, 2009, Slussareff *et al.*, 2016; Rocha, 2014). The literature indicates that most evaluations and assessments in SGs occur at the reaction level of Kirkpatrick's model (Kirkpatrick; Kirkpatrick, 2006), with the collection of student satisfaction, motivation and other perceptions. At the same time, the assessment of the learning level, which measures the change in attitudes, knowledge and skills, is neglected (Salas *et al.*, 2009; Savi *et al.*, 2010; Rocha *et al.*, 2015; Oliveira *et al.*, 2018). This issue is because of the lack of approaches (processes, models, frameworks) that address the planning of this assessment during the SGs development lifecycle, as there is a greater concern in game design and not in the design of the student's performance assessment (Rocha, 2014; Oliveira *et al.*, 2018; Emmerich; Bockholt, 2016).

It is important to collect in-game data to measure learning and provide results reports to stakeholders (teacher/instructor, institution, among others) (Salas *et al.*, 2009). The in-game assessment offers the opportunity to use the game itself and employ alternative and less obvious forms of assessment that could (and should) become a game element, such as a Stealth Assessment (Bente; Breuer, 2009). Continuous and immediate feedback is also important because it supports the effective learning of pedagogical goals (Trybus, 2010). A model that aggregates these aspects is relevant to the efficiency and effectiveness of SGs (Rocha, 2014; Trybus, 2009; Chaudy; Connolly, 2019; Emmerich; Bockholt, 2016).

This paper is an extended version of the Oliveira & Rocha (2020b), which presented AvaliaJS, a conceptual model for planning the design and execution of assessment in SGs. We aim to describe an evaluation of the model by an expert panel. AvaliaJS was conceived through a holistic view on the evaluation and assessment approaches (methods, methodologies, processes, frameworks, models), in the context of SGs, from a literature review. The expert panel is a method used to gain expert knowledge of a particular domain or area (Beecham *et al.*, 2005). It was used to conduct a technical evaluation of the AvaliaJS model and obtain a series of recommendations. The evaluation of the AvaliaJS by the expert panel was structured according to the GQM method (Basili *et al*., 1994), which specified the goals, questions and metrics of analysis. A questionnaire was used to collect data from the experts' answers, after analyzing the AvaliaJS model.

This paper is organized as follows: Section 2 presents the theoretical background, with the main concepts involved, such as serious games, game evaluation and assessment, and data collection instruments. Section 3 presents the related works and Section 4 describes the proposed conceptual model (together with the canvas model and the assessment project document). Section 5 presents the evaluations and results, with examples of AvaliaJS model used for planning three serious games and the evaluation based on the expert

perspective. Section 6 discusses the results and Section 7, the final considerations.

2 Background

This section summarizes the concepts, theories, techniques and instruments that support the construction of the conceptual model proposed and evaluated in this paper.

2.1 Serious Games

A game "is a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome." (Salen; Zimmerman, 2003, p. 80); the term "serious" points out that they are games aimed at purposes other than pure entertainment (Abt, 1987).

Serious Games (SGs) are effective for teaching and training students of different ages, mainly by four factors: (1) are highly motivating; (2) provide efficient communication about pedagogical concepts and contents; (3) provide a contextualized representation of the problem to be taught; (4) make students take on realistic roles, such as tackling problems, formulating strategies, making decisions, and getting immediate feedback on the consequences of their actions (Alvarez; Djaouti, 2012; Salas *et al*., 2009; Rocha, 2014; Abt, 1987).

SGs can be digital (computer use) or analog games (boards and physical objects), and grouped into genres, such as adventure, action, puzzle, strategies (Abt, 1987; Shell, 2008; Petri, 2018). In this context, Djaouti *et al*. (2011) have developed a model for serious game classification, consisting of three aspects: (1) ***gameplay***: intended to provide information about the game structure of the SG, can be *game-based* (well-defined rules, such as *Mario World*) ou *play-based* (does not feature stated goals, such as *Sim City*); (2) ***purpose***: indicates the overall goal of the game, which is divided into three types: (a) message-broadcasting (educative, informative, persuasive and subjective); (b) training (to improve cognitive performance or motor skills); (c) data exchange (collecting player information); and (3) ***escope***: indicates the kind of market (health, military, education, religious) and the audience (general public, professionals and students).

2.2 Assessment, Evaluation and Performance in Serious Games

2.2.1 Conceptualization and Types of Assessment and Evaluation

There are two different processes to help students build learning competencies: evaluation and assessment (Baehr, 2010). The *evaluation* is defined as the process of passing judgment about learner performance or SG effectiveness (usability aspects), based on defined criteria and evidence, aiming to reinforce, guide and correct behavior of the evaluated in their tasks or improve the SG effectiveness (Daoudi *et al*., 2017). The *assessment* is defined as the process of collecting, reviewing and using data, to improve the student's current performance (help to learning and skills development), to provide them feedback on their errors and hits (Bellotti *et al*., 2013; Daoudi *et al*., 2017).

The main objective of the assessment is to provide feedback to stakeholders, which may include students, teachers/instructors, and coordinators/managers (Zinovieff; Rotem, 2008). Student performance refers to increased student knowledge and capacity as a result of learning activity (Ariffin *et al*., 2014). The performance assessment determines the degree to which the student applies in the real world the competencies acquired (Salas *et al*., 2009). Competence is "the ability of an individual to perform a specific activity or work with quality" (Durand, 2000). The competence is divided into three dimensions: (1) **knowledge**: refers to a set of information stored in the person's memory, i.e., having information; (2) **skills**: refers to the ability to make productive use of knowledge, i.e., it is to have the technique and ability to apply knowledge; and (3) **attitude**: refers to the person's predisposition (wanting to make/determination) to work, objects or situations, i.e., applying the skill (Savi et al., 2010; Rocha, 2014).

The evaluation and assessment can occur at different times of the teaching-learning process (Hettiarachchi et al., 2013): (1) **diagnostic**: carried out in the beginning, to analyze the student's previous knowledge; (2) **formative**: carried out during the process, to improve and develop the competencies; and (3) **summative**: carried out after the learning period, aiming to measure and classify the student progression. Self-assessment can be used to reflect the student's evolution (Rocha, 2014).

The assessment can be of the **reaction** ("did the students like the game?"), **learning** ("did students learn from the game?"), **behavior** ("are students applying the new knowledge?") or **results** ("did the game have an impact on the results?"), according to Kirkpatrick's model (Kirkpatrick; Kirkpatrick, 2006; Kirkpatrick; Kirkpatrick, 2016). At the reaction level, the student's perceptions are evaluated in three dimensions: **motivation** (relevance, confidence, satisfaction and attention), **engagement** and **self-assessment**. At the learning level, changes in what the user knows about the content after the activity are identified. The learning assessment is related to the measurement of knowledge, skill, attitude and commitment (Kirkpatrick; Kirkpatrick, 2016).

Feedback is information provided by an agent regarding aspects of performance or understanding and complement evaluation (Hattie; Timperley, 2007). In serious games, feedback can be used to report **game progress**, **learning feedback** and **user interaction** (Ifenthaler *et al*., 2012; Rocha, 2014; Chaudy; Connolly, 2019).

In the context of SGs, the assessment can be classified according to its purpose (Ifenthaler *et al*., 2012): (1) ***external***: data collection processes that use tools external to the game, such as observation, tests, debriefing, interview; (2) ***internal***: tools and techniques applied within the game, such as log-file, monitoring of states, learning analytics; (3) ***before***: occurs previous to interaction with the game; (4) ***during***: occurs when the player interacts with the game; and (5) ***after***: occurs when the player finishes his/her interaction with the game.

Still according to Ifenthaler *et al.* (2012), when it comes to internal evaluation, the types of records of players' actions can be classified: (i) ***game score record***: refers to the scoring methods and time needed for completing a specific task; and (ii) ***interaction record***: describes student behavior during the game, collected through log-files or clickstreams,

information trails, monitoring of status, among others (Ifenthaler *et al*., 2012). Based on Rocha (2014), this paper added the ***performance record***, which refers to recording data related to the learning of the game content (player competencies), e.g. errors and hits, sequences of actions, i.e., it is the player's performance with the taught content.

2.2.2 Theories of learning, training, reaction and feedback

Some theories are applied in SGs to support the planning and requirements gathering, in the context of learning and training, reaction and feedback: (1) **theories about learning and training**: refers to pedagogical theories about learning and training, used to describe the motives and processes of learning and human performance, events and teaching methods. For example, Bloom's taxonomy (Bloom, 1956), training program evaluation theory (Kirkpatrick; Kirkpatrick, 2016), learning-experiential theory and learning style (Kolb; Kolb, 2005), principles of affective learning (Trybus, 2010); (2) **theories about user's reaction:** refers to theories related to player satisfaction, e.g., ARCS (Attention, Relevance, Confidence and Satisfaction) model and expectation-value theory of motivational strategies (Keller, 1987; 2009), flow theory (Csikszentmihalyi, 1990), engagement (Boyle *et al*., 2012); and (3) **theories about feedback**: refers to theories that define the types, how and when to make and expose feedback, as well as support in the description and classification of feedback. For example, individual performance theory (Salas *et al*., 2009), feedback dimensions (Bee; Bee, 2000 *apud* Rocha, 2014), feedback classification and levels (Hattie; Timperley, 2007).

2.2.3 Data Collection Techniques and Instruments

In the context of SGs, different techniques and data collection instruments can be used in: (1) **external assessment:** questionnaires (e.g., reaction and self-assessment), interview, pre-test and post-test, observation, focus group, debriefing, think-aloud protocol, conceptual map, chat, forum, among others (Rocha *et al.*, 2015; Chaudy; Connolly, 2019; Ifenthaler *et al*., 2012; Eseryel *et al.*, 2011; Oliveira *et al*., 2019); and (2) **internal assessment:** phases of the game with data collection of user's actions (records of score, interaction and performance), phases with a questionnaire (pre- and post-test, reaction, self-assessment and player profile questionnaires), player registration, learning analytics, among others (Rocha, 2014; Chaudy; Connolly, 2019; Ifenthaler *et al.*, 2012; Eseryel *et al.*, 2011).

3 Related Works

Studies on assessment in serious games can be classified into three groups: (i) **approaches (methods, methodologies, processes, frameworks, models) to SG/game development and assessment/evaluation** (Rocha, 2014; Pereira Junior; Menezes, 2015; Chaudy; Connolly, 2019; Ibrahim; Jaafar, 2009; Kiili, 2005; Westera *et al*., 2008; Victal; Menezes, 2015; Akilli; Cagiltay, 2006; Rocha *et al*., 2017; Yedri *et al.*, 2018; Jappur *et al.*, 2014; Petri, 2018); (ii) **canvas template for game design** (Sarinho, 2017; Sousa, 2014; Carey, 2015; Star *et al*., 2016; Kornonean *et al*., 2017; Walker, 2015); and (iii) **systematic reviews and mappings that analyze evaluation/assessment in SGs and educational games** (Battistella *et al*., 2014; Petri; Wangenheim, 2016; Lopes *et al*., 2013; Petri; Wangenheim, 2017; Wangenheim *et al*., 2009; Calderón; Ruiz, 2015; Abdulmajed *et al*., 2015; Wang *et al*., 2016; Alfarah *et al.*, 2010; Daoudi *et al*., 2017).

In group (i), most game development approaches do not specify assessment planning in the development life cycle. They are more concerned with collecting player action data, such as the number of errors and hits, and there is no focus on planning that involves the different types of assessment and evaluation. The approaches that specify assessment include various types of evaluation/assessment, such as formative, diagnostic and summative. However, some approaches do not include external evaluation in the evaluation/assessment planning; or do not specify the feedback during the game; or are limited to the created game used to evaluate the approach. For example, Petri (2018) presents MEEGA+, a model that contains the planning and conducting process to evaluate the quality of educational games (in terms of player experience and usability). However, it does not specify or support an in-game assessment, such as registration, interaction and performance data collection.

In the group (ii), the canvas models do not specify the student performance assessment and evaluation of the game itself. Just the work of Star et al. (2016) supports student performance assessment planning and game evaluation, however, it is specific to prosocial games.

In group (iii), almost all analyzed games are developed disregarding the planning of students' performance assessment. The games are evaluated by collecting players' opinions (reaction - level 1). The learning assessment (level 2) is usually performed through pre- and post-tests, without assessments during the game (formative evaluation and data collection of players' actions during the game).

The studies analyzed in the three groups show the importance of internal and external assessment in SGs. External assessment alone (not aligned with the internal assessment) may neglect important changes during the learning process, as it is mainly focused on the application of pre- and post-tests (Eseryel *et al.*, 2011). This approach can make it difficult to provide immediate feedback regarding the content covered in the game. The internal assessment, focused on in-game data collection, if well designed, can provide detailed information about learning processes as well as immediate and personalized feedback for the student (Ifenthaler *et al.*, 2012; Eseryel *et al.*, 2011). Thus, internal and external assessments should be strategically designed to ensure student learning through SGs, both before, after and during interaction with the game.

In-game data collection and assessment can provide constant and personalized feedback to the players (professionals and students) while running the game (Chaudy; Connolly, 2019). In addition, they are also important for measuring learning and offering to report for stakeholders (e.g. teachers and institutions) (Salas *et al*., 2009) and evaluating the game effectiveness (final product), after the game (Hays, 2005). Thus, a model that allows the planning of the design and execution of the assessment is relevant to support the evaluation and development team, as well as to document and analyze the results of this process and final product.

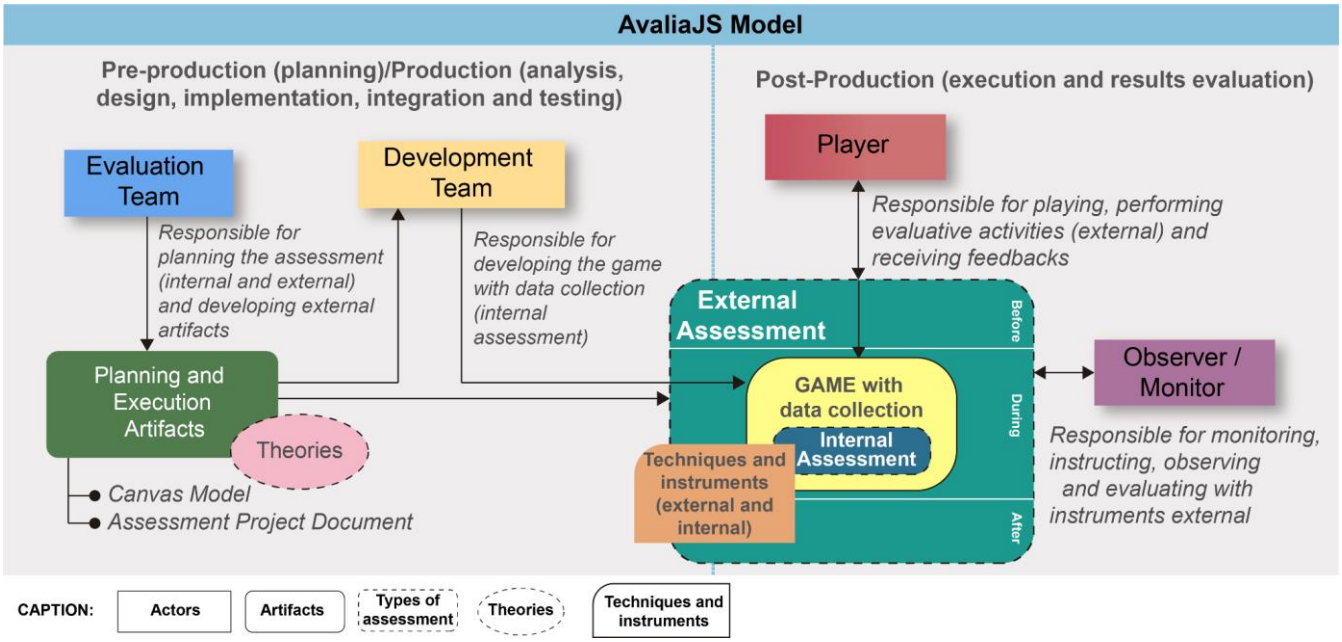

Figure 1. AvaliaJS overview: a conceptual model for planning of the design and execution of student performance assessment in SGs.

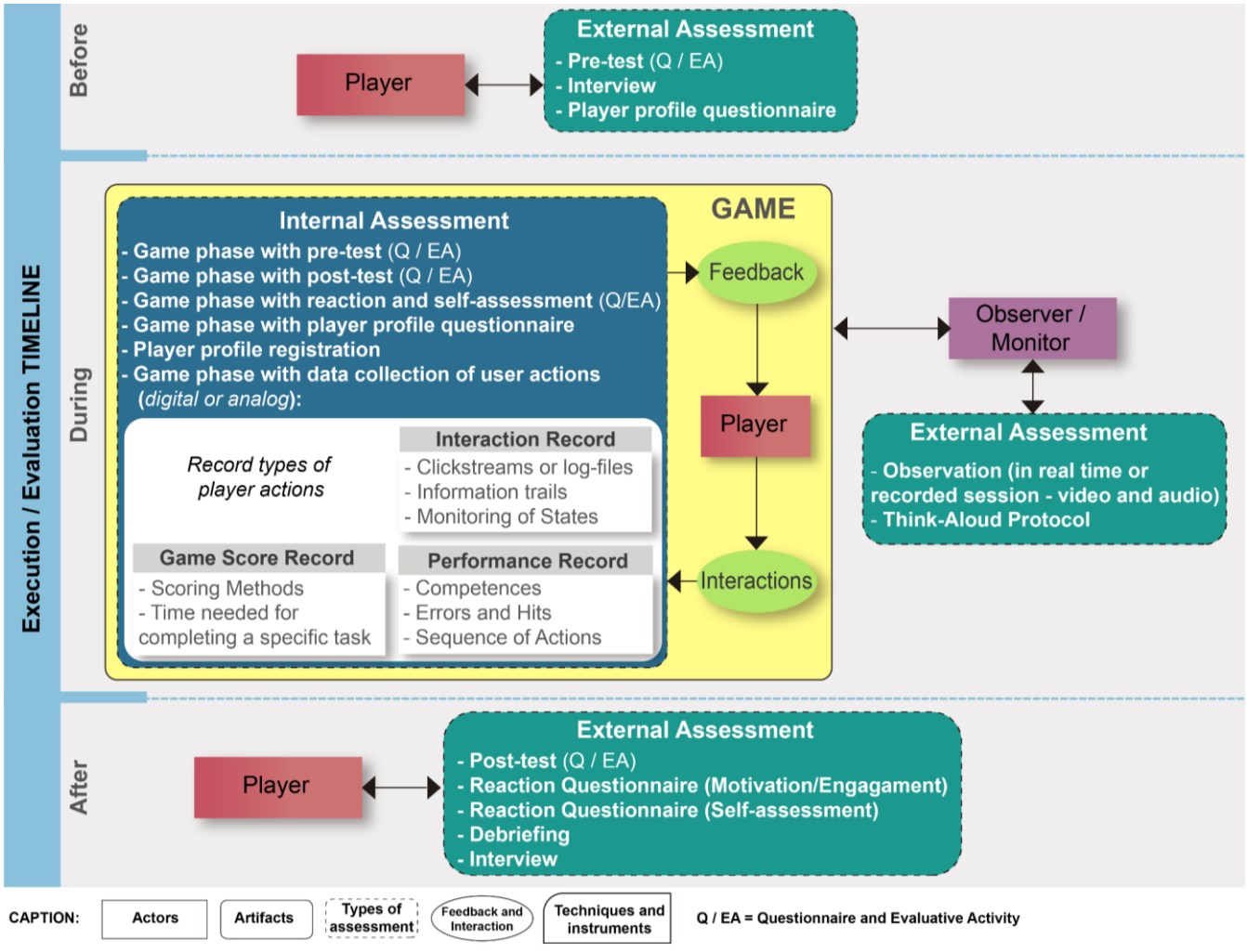

Figure 2. Detailed overview of the execution/evaluation step of the AvaliaJS model.

4 AvaliaJS Model

The conceptual model proposed was called **AvaliaJS** (in Portuguese: *Avalia= Avaliação (Assessment) + JS= Jogo Sério (Serious Games)).* It was conceived through a holistic view on the evaluation and assessment approaches (methods, methodologies, processes, frameworks, models), in the context of SGs, from a literature review. AvaliaJS aims to develop games considering the planning of the design and execution of the student performance assessment in SGs and immediate and constant feedback.

In AvaliaJS model is considered the development life cycle of an SGs, based on Rocha (2014), that is formed by three stages: (1) pre-production (initial planning), (2) production (analysis, design, implementation, integration and testing) and (3) post-production (execution/evaluation and analysis of results). Thus, the AvaliaJS is divided between the pre-production/production and post-production stages as shown in **Figure 1**.

In general, the evaluation team is responsible for (1) planning the external and internal assessment, using the appropriate theories according to the focus of the SGs; and (2) developing artifacts for external and internal assessment, such as questionnaires, observation checklist, interview protocol. The internal assessment artifacts, with the description of techniques and decisions for internal assessment, generated by the evaluation team, are delivered to the development team, which develops the game with data collection (internal assessment). The external assessment artifacts are designed to be used by the player before, during, or after the game application. Thus, the player plays, performs assessment (external) activities and receives feedback; and the observer/monitor monitors, instructs, observes and evaluates with external instruments. **Figure 2** illustrates a detailed cutout of the execution/evaluation stage, in which some techniques, assessment instruments and data collection (internal and external to the game) are presented.

The AvaliaJS has artifacts for the assessment planning, which aim at the description of the theories, techniques and instruments, both external and internal, which will be used for the execution of the assessment: (1) canvas model (high-level planning), as shown in **Figure 3**, and (2) assessment project document (description in low level). The *canvas model*[1] allows quick identification of the elements and activities required for the planning of design and execution of the student's performance assessment in SGs. In the *assessment project document*[2], what was planned in the canvas model should be detailed.

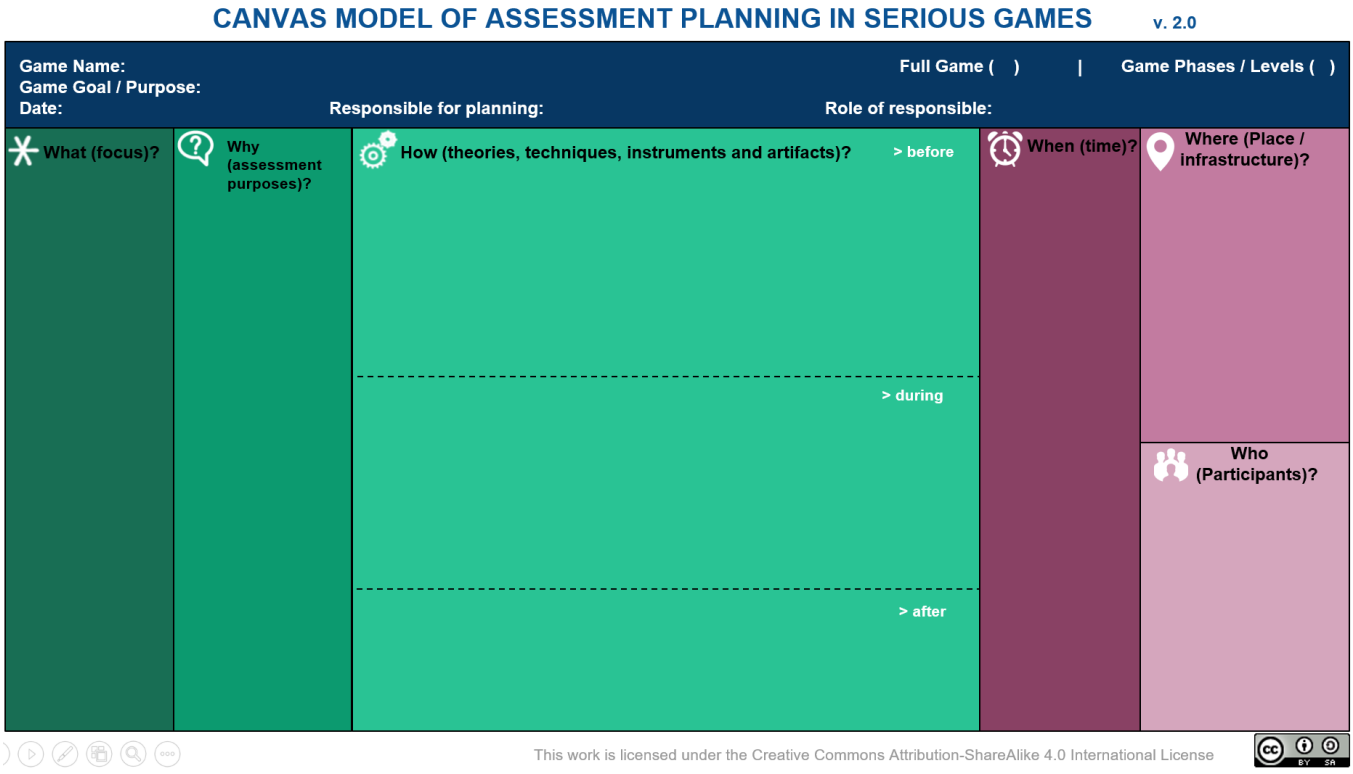

Figure 3. Canvas model for the planning of assessment in SG.

The proposed canvas model (as illustrated in **Figure 3**) was developed based on the **5W2H** method (Rossato, 1996) and questions that can be used to plan an assessment in the classroom context in higher education (Falchikov, 2005). The 5W2H method is an action plan that allows elements and tasks to be identified quickly, during the development life cycle. **Table 1** presents the relationship among the questions of Rossato (1996), of Falchikov (2005) and the proposed canvas model.

[1] http://www.bit.ly/CanvasModelAvaliaJS last access on 04/11/2021.

[2] http://www.bit.ly/ProjDocAvaliaJS last access on 04/11/2021.

Table 1. Relationship among the questions for assessment planning.

5W2H (1996)	Falchikov (2005)	AvaliaJS Canvas model
What (what will be done?)	What will be the assessment focus? (what to assess?)	***What*** will be assessed during the game? (*focus*)
Why (why will it be done?)	The assessment purpose (why to assess?)	***Why*** should the assessment be carried out? (assessment *purposes*)
Where (where will it be done?)	-	***Where*** will each assessment be carried out? (*mode, place, context and equipment*)
When *(when will it be done?)*	*When* should each assessment be carried out? (w*hen to assess?)*	***When*** *should each assessment be carried out?* (*time/duration/deadline*)
Who (by whom will it be done?)	For *whom* the results will be provided (*who assesses?)*	***Who*** *is involved in carrying out the assessments?* (*participants/role*)
How (how will it be done?)	What evaluation methodology will be used? (how to assess*?*)	***How*** *should the assessment be carried out?* (*theories, techniques, instructions and artifacts)*
How much (how much will it cost?)	-	-
-	How reliable or valid are student assessments? (*how well do we assess?*)	-
-	What will be done with the results of the evaluation? (*Whither? What next)*	-

The questions selected for the development of the canvas model allow quick identification of the elements and activities necessary for the planning of the design and execution of the student performance assessment in SGs. Falchikov (2005) did not address the questions of where they will be carried out and how much the assessments will cost. However, he included the reliability and validity of the assessments, as well as the planning of what will be done with the results. The cost was not considered, because AvaliaJS focused on how to do the planning (goals, techniques and execution), and also the analyzed games did not address the cost. In addition, the reliability was not added to the model, because the techniques and instruments listed to be used in the assessment were consulted and grouped in the model from the literature review, which increases reliability. Another question not addressed is the planning of what will be done with the results. It was not inserted because assessment analysis and results were not discussed in using the AvaliaJS model. However, the three questions can be considered in future work.

The canvas model was developed incorporating the characteristics for the construction of a Business Model Canvas (BMC), presented by Osterwalder (2004). BMC is a tool that gives a preformatted overview and allows you to develop and sketch new or existing business models, through blocks (Osterwalder, 2004). It is important to reinforce that some canvas models for game design were inspired by BMC, such as *Unified Game Canvas* (*UDC*) (Sarinho, 2017). In this paper, the following characteristics were considered: (1) *organization by influence*: it organizes the sections that have the greatest influence on each other; (2) *grouping by relation*: the set of sections that have a relationship with each other or offer a specific scenario; and (3) *atomic meaning of the sections*: each section deals specifically with a single subject (Osterwalder, 2004).

Thus, the sections of the canvas template were organized to deal with a specific subject and influence each other. The sections were grouped by color, being differentiated by the target of the assessment planning: (1) *blue*: refers to the contextualization of the game; (2) g*reen*: refers to the project planning of internal and external assessment, such as focus, objective and theories, methods, instruments and artifacts; and (3) *purple*: refers to the planning of the assessment execution, such as time, place and participants.

Figure 4 presents the canvas model filled with the main contents for assessment planning. The canvas model does not have a defined fill order. However, it is preferable to start with the blue section: fill in the game name and overall objective fill date, names of the responsible and their function, mark whether the assessment planning is of the full game or a specific phase/levels (e.g., the prototype of only one game mission). Then, the green and purple group blocks can be filled in parallel, according to the project needs.

Figure 5 presents the summary of the main content used for planning the design and execution of the assessment in SGs. The contents were added according to the theories, techniques, instruments and artifacts found in the literature (in red, features not addressed by the model).

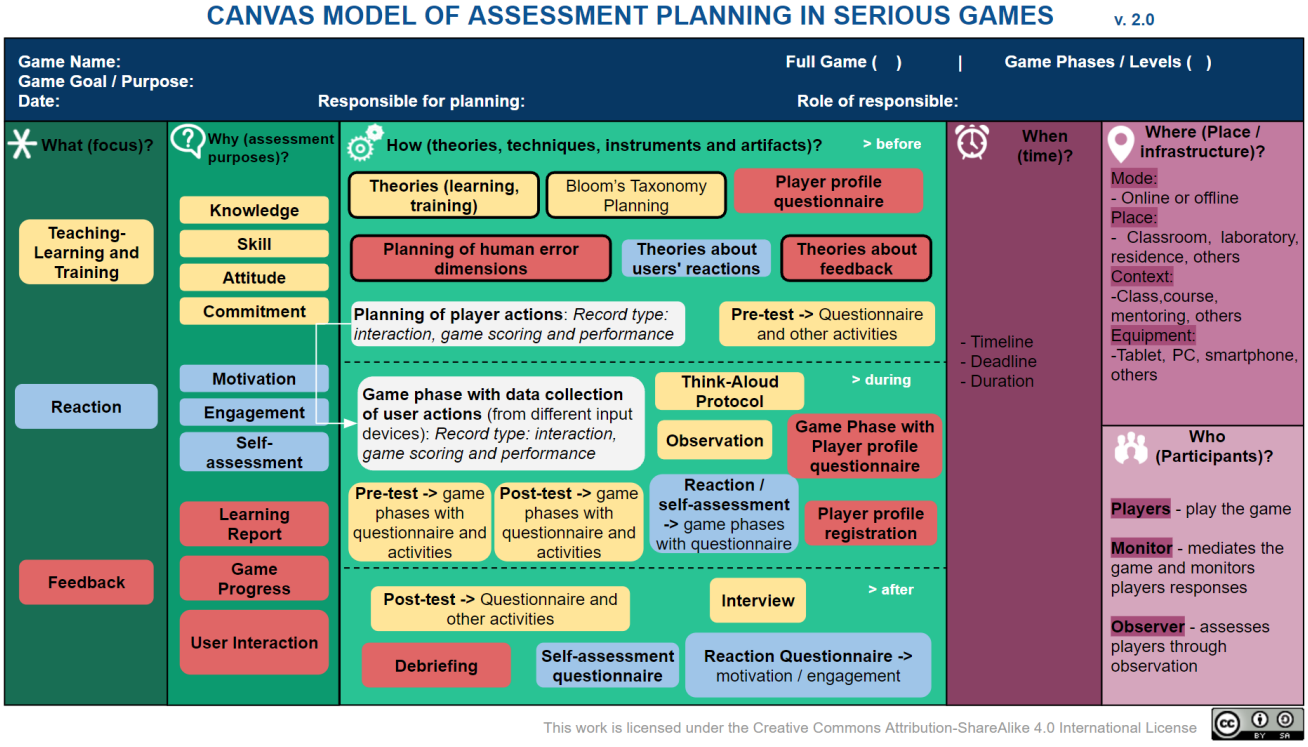

Figure 4. Canvas model for the planning of the assessment in SGs: examples of contents.

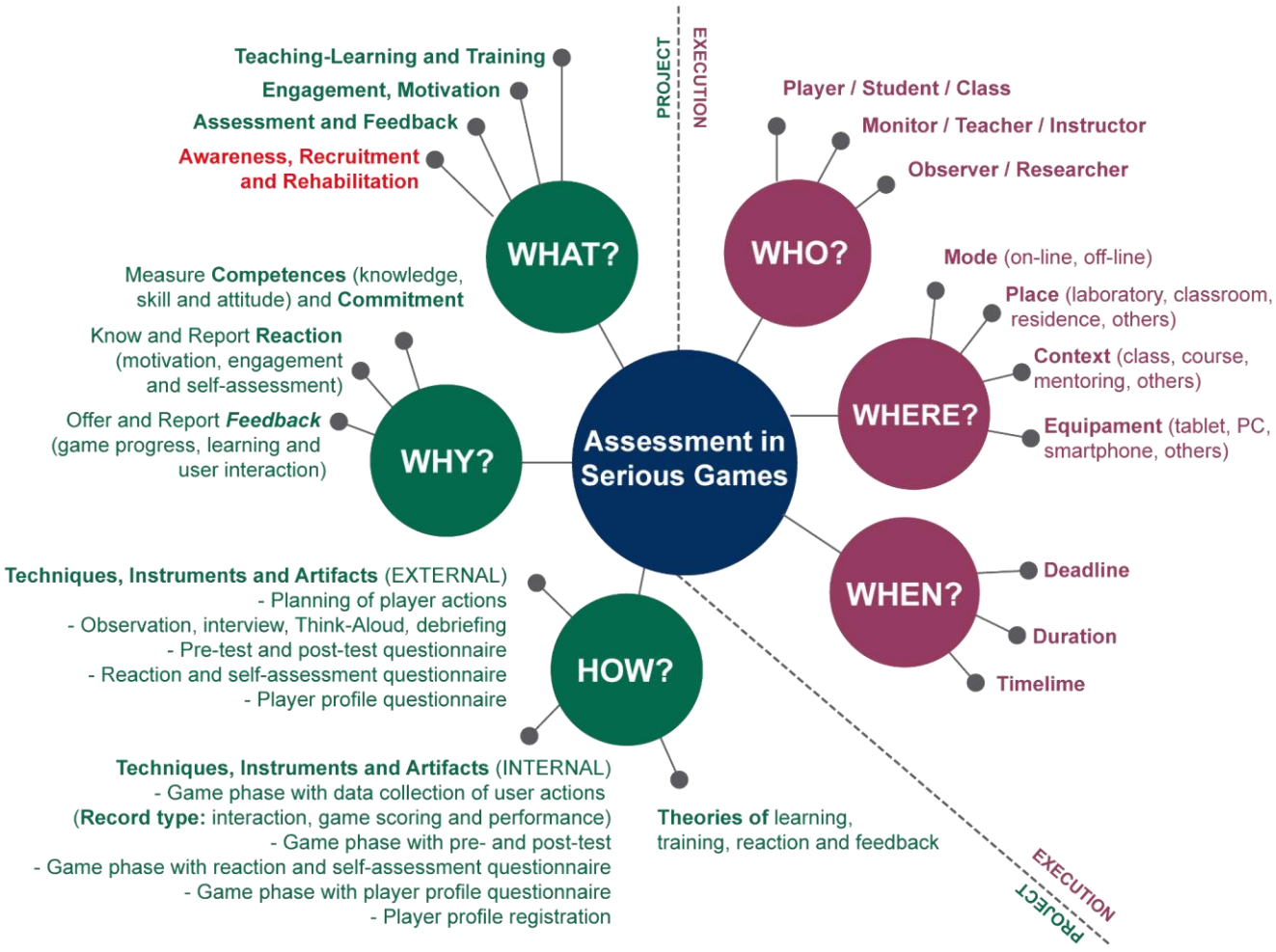

Figure 5. Summarized diagram of the canvas model content (in red, features not addressed by the model).

5 Evaluations and Results

This section presents (1) the results of the application of three serious games, to exemplify the use of the AvaliaJS model as proof of concept; and (2) the evaluation of the AvaliaJS model through an expert panel.

5.1 Application and exemplification of the AvaliaJS

The objective of the application and exemplification of the AvaliaJS is (according to Goal/Question/Metric (GQM) method, proposed by Basili *et al.* (1994)) :

> **Analyze** the games *GLPSobControle (G1), Guerra em Alto Mar (G2)* and *Expedição Antártica (G3)* **for the purpose of** evaluating and exemplify the use of the conceptual model as proof of concept **with respect to** coverage of contents to fill in the model **from the viewpoint of** researchers **in the context of** development and assessment/evaluation in serious games.

Filling out the canvas model and the project document was prepared by the researchers (who also participated in the productions of these games (G1: author 2, G2: author 1, G3: authors 1, 2 and 3)), after defining the games analyzed. To conduct the evaluation, researchers should answer the following question: *does the model include and cover the content necessary for planning and evaluating the player assessment in the serious game (as performed in each game)?* As this is a qualitative evaluation, the metrics were defined to identify the contents to fill the model: focuses, purposes, theories, techniques, instruments, artifacts and infrastructure. These games were chosen because they are focused on teaching-learning, training, evaluation, motivation and engagement and because they are available (game/source code and documentation). The application (use and example) of the canvas model and the project document in these three different games, and with different pedagogical objectives, allows an evaluation and analysis of AvaliaJS. We present below the results of the interpretation according to the GQM method.

The games *GLPSobControle (G1), Guerra em Alto Mar (G2)* and *Expedição Antártica (G3)* were used to exemplify the canvas model and the assessment project document. The *GLPSobControle* (in English: Liquefied Petroleum Gas Under Control) is a digital game for training and assessment of firefighters on the control of kitchen gas leaks (Rocha, 2014). The *Guerra em Alto Mar* (in English - War on the High Seas) is a board game for motivating and engaging students in learning Python programming language (Oliveira; Rocha, 2019). The *Expedição Antártica* (in English - Antarctica Expedition) is a digital game for teaching-learning about the knowledge of citizen science set in Antarctica (Oliveira *et al*., 2019). A comparison of the analyzed games is presented in **Table 2**.

The filling the canvas model and the assessment project document, in the three games, began with contextualization. The games were situated in a specific context, describing the name, gameplay and genre, general objective (purpose), target audience, the role of the person responsible for assessment planning and a summary of the main objectives of the game (full game or phases/missions). Then, the scope of the assessment project was defined, such as the focus (what?), objective (why?) and theories, techniques and instruments (how?), followed by the description of the execution part of the assessment, focused on when and where the assessment occurs and who is involved in it. The canvas templates and assessment project documents of the exemplified games can be downloaded from the online repository[3].

Table 2. Comparison of the analyzed games.

Info \ Game	G1	G2	G3
Type of game	3D/2D	analogic	2D
Genre	simulation	board	*RPG*
Phases/Levels (analyzed)	7 phases	1 phase with rounds until you have a winner	1 phase with 4 mini-games
Target audience	firefighters	computer students	undergraduate students
Motivation/ Domain	kitchen gas leak training	motivation for learning Python	citizen science (whale mission)

About the objective of the assessment ("*why?*"), based on our filled documents, we can observe that the model describes the assessment objective of the three SGs analyzed. The assessment objectives were divided according to the focus of the game: learning and training (knowledge, skill, attitude and commitment), reaction (motivation, engagement and self-assessment) and assessment and feedback (learning report, game progress and user interaction). The games *GLPSobControle* and *Expedição Antártica* present similar objectives, differing that in the first there is self-assessment and in the second the commitment. At *Guerra em Alto Mar*, the objectives of the assessment are focused on promoting motivation and engagement, measuring knowledge and reporting learning and game progress.

About the theories, techniques, instruments and artifacts used, the two digital games were developed so that their phases could collect players' actions and record performance, interaction and score game. They take into account internal and external data collection, to implement the evaluation and assessment of training, teaching-learning and reaction. At *GLPSobControle*, the first phase is a 2D card game to check what the learner already knows about what will be trained, which is characterized as a phase with pre-test (Rocha, 2014). The following phases were implemented in 3D to record the sequences of actions of the learner. Such sequences produce results that give feedback on the player's performance at the end of the training, such as a wrong sequence that generates an explosion, or a correct sequence, which indicates the success of the training. The last phase is a questionnaire of self-assessment and evaluation/assessment of the training program, that collects the user reaction.

The game *Expedição Antártica* has no linear phases, as in *GLPSobControle*, and the player can start with the mini-

[3] https://bit.ly/LGPCanvas, https://bit.ly/AntarcticaCanvas, https://bit.ly/WarCanvas, last access on 04/11/2021.

game they prefer (only the final mini-game that, to be released, needs all others to be completed) (Oliveira *et al.*, 2019). Because it is an RPG game, the game score record is designed to sustain the player's experience and life bar, which increases or decreases according to players' actions, such as chatting with a mentor, playing a mini-game, earning reward, completing or redoing missions. In addition, in some mini-games, the player receives tips from mentors to advance in their activities. Also, in the *Expedição Antártica*, the database attributes were divided into game score, performance and user interaction records, which helps in planning the data to be collected. About external instruments, pre- and post-test, player profile and reaction questionnaires were used. During the interaction with the game, the players were able to verbalize their thoughts through the Think-Aloud protocol, being observed by the researchers (Oliveira *et al.*, 2019). At the end of the interaction, an interview and a debriefing session were held.

In *Guerra em Alto Mar*, because it is an analog game (board), the player actions are recorded through a result collection form, focused on the performance and game score record. A monitor manually recorded the results of the answers (if they missed or got it right) of the players to the quiz and the response time of each (Oliveira; Rocha, 2019). As external instruments, the students were observed by the researchers and, after interaction with the game, answered a reaction questionnaire. This game was used to exemplify that the model proposed in this paper covers the assessment planning in analog games, however, it needs to be validated in the creation of other analog games.

In general, the model allows the inclusion of artifacts external to the game (such as questionnaires and interviews) and internal instruments, which aim to collect players' data, through the records of game score, performance and user interaction. Thus, the model allows the support of the use of phases with pre-test, self-assessment and satisfaction questionnaires to the game, such as the case of *GLPSobControle*.

The planning and documentation of feedback, in the use of theories and dimensions of human errors, allowed a detailed view of pedagogical aspects to be assessed in the games. The constant and immediate feedback is perceived, in digital games, through punctuation, visual and sound feedback of actions and consequences attributed to errors and hits of players, such as receiving a camera as a reward, in the *Expedição Antártica*, or an explosion when generating static ignition, in the game *GLPSobControle* (Rocha, 2014; Oliveira *et al*., 2019). The game *Expedição Antártica* has mentors who give tips in response to some wrong actions of players in mini-games and dialogues. In *Guerra em Alto Mar*, to advance in gameplay, the student needs to answer a quiz. The right or wrong feedback is given through the "Answer Letter", which contains the correct alternative and an explanation of the quiz. About the final feedback, in the case of the game *GLPSobControle*, the data of the result of the apprentice's performance are presented at the end of the training, on a screen that shows, for example, the number of victims, duration of the training, results of phases, number of corrections (Rocha, 2014). The game *Expedição Antártica* did not implement a report at the end of the game, because of the lack of time and the project scope (Oliveira *et al*., 2019). The *Guerra em Alto Mar* used the results sheet as a report to discuss the errors and correct answers of the students in each question (Oliveira; Rocha, 2019).

Regarding the execution planning, the model provided the schedule planning and duration of the assessment activities (external and internal). In the case of the exemplified games, the evaluation was performed in just one day (Rocha, 2014; Oliveira *et al*., 2019; Oliveira; Rocha, 2019). However, it may happen that, for example, a pre-test questionnaire is applied one week before the game is applied. What sets the time is the timeline of the game project to be developed. In addition, the model also provided the planning of the place/infrastructure and participants involved in the evaluation/assessment session.

5.2 Evaluation of the AvaliaJS by Experts

The evaluation of the AvaliaJS model was carried out through an expert panel (Beecham *et al.*, 2005). The evaluation planning of the AvaliaJS model was structured according to the GQM method (Basili *et al.*, 1994). Thus, the objective of the study was defined as:

> **Analyze the** AvaliaJS model (diagrams + canvas template + project document) **for the purpose of** evaluating the quality **with respect to** its correctness, consistency, understandability, unambiguousness, completeness, authenticity, flexibility and usability **from the viewpoint of** experts in serious games and assessment involved in the production and development of the games used to apply and exemplify AvaliaJS model **in the context of** student performance assessment in serious games.

In this paper, quality is defined as the degree and/or capacity of a model to meet the needs, expectations or requirements specified by the user for a specific goal (IEEE, 2010). To collect the data, a questionnaire was created and used, reported in Oliveira (2020), whose metrics and questions were based on the works of Savi et al. (2010), Rocha (2014) and Petri (2018). The validation of the items of the questionnaire was performed through a semantic analysis (Zerbini *et al*., 2012) and a checklist for writing and evaluating items, adapted from Coelho *et al.* (2020) and Mourão & Meneses (2012). The questionnaire can be viewed in **Appendix B**.

The questionnaire was divided and organized into three parts: (1) collection of the expert's profile; (2) items related to quality characteristics; and (3) additional open items, to collect insights from the main strengths and weaknesses of the evaluated model. A protocol was developed to guide the experts in the evaluation of the AvaliaJS model.

To evaluate AvaliaJS (diagrams + canvas model + project document), six experts in the area of serious game development and evaluation were invited, by sending the email with the evaluation protocol. The evaluators were selected for convenience because they are researchers involved in the development of the serious games used to exemplify the AvaliaJS model and have knowledge of human performance assessment in different perspectives and areas (training, education, psychology, neuroscience, gaming and computing). For the game *GLPSobControle*, a domain expert was invited. In the game *Guerra em Alto Mar*, an expert teacher was invited. For the game *Expedição Antártica*, four professionals were invited: one expert-teacher of the accessibility team and another from the evaluation team, a

developer-student of the development team and a student of the content team. The evaluators were instructed to evaluate both the artifacts of the game that they helped develop and the artifacts of the other games. Of the six specialists, three are doctors, one is a doctoral student, one is a master's student and one is an undergraduate student. Four evaluators are in the area of computer science, one has a degree in psychology and the other in the area of neuroscience.

The **first part of the questionnaire** aimed to collect information about the experts' experience related to the development, use and evaluation and assessment of SGs. **Figure 6** illustrates that most experts have developed (66.6%), used (50%) or evaluated (66.7%) less than five games.

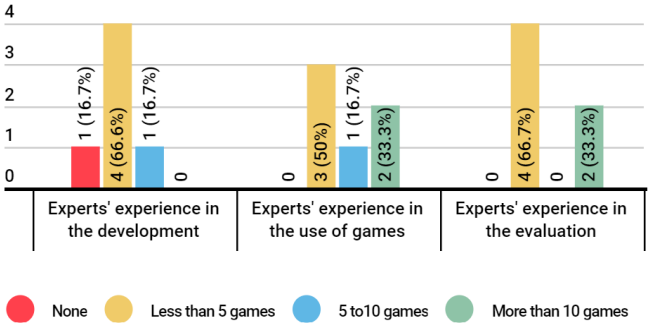

Figure 6. Experts' experience in the development, use and evaluation of serious games.

In the case of the development experience, probably the expert who scored who developed "none game" must have considered that he/she did not because he/she did not participate in the design and programming of the game. Regarding the use, there was no question whether the use was for personal purposes or to be applied in the classroom context (in workshops, classes, short courses and related). Regarding the evaluation, the experts were asked how they assess student performance in SGs. **Figure 7** illustrates that 83.3% of the experts use the pre- and post-test technique to collect information about the players' knowledge and 66.7% use interviews/debriefing at the end of the game interaction; 16.7% self-assessment and another 16.7% use observation. The competencies assessment, which involves knowledge, skill and attitude is the objective of assessment most used by experts, with 66.7%; commitment assessment is not used and 16.7% of experts assess motivation. In addition, one of the experts described the use of other methods for evaluation and assessment, such as the measure of time, the measure of look (number and time of fixings and tracing), hit and error in each phase, automatic in-game registration and the production of the report with graphs and tables that the game produces. One of the experts said he/she doesn't usually evaluate. Such response may be incoherent since all experts stated that they evaluated at least one SG, as shown in **Figure 6**. The authors understand that, probably, the expert considered "I don't usually evaluate" because he/she did not participate in the evaluation process with the students, interpreting the evaluation of the SG in aspects of usability and not of student's performance assessment.

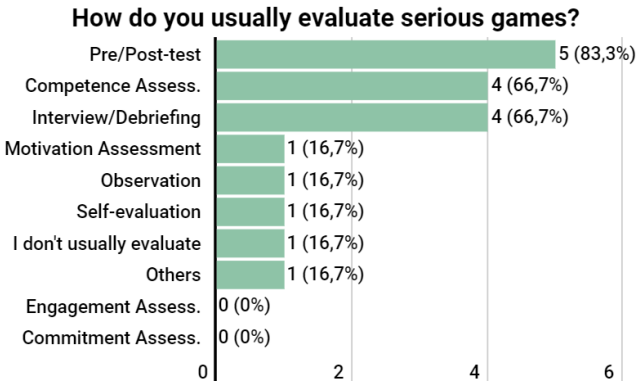

Figure 7. Types of assessment/evaluation objectives and techniques used by experts in SGs.

The experts were asked about their knowledge of other planning artifacts of the design and execution of the student performance assessment in SGs. Most evaluators (83.3%) commented that he/she did not know other artifacts with a proposal similar to that presented in this paper. In addition, experts were asked to comment on the degree of difficulty in planning the assessment of reaction and learning, inside and outside SGs. In general, experts are aware of the complexity of assessment planning in SGs and the difficulty in finding tools that support this, which aligns with the thoughts of several authors, such as Emmerich *et al.* (2016), Rocha (2017) and Oliveira *et al.* (2018).

The **second part of the questionnaire** consists of five items with dichotomous scale (yes or no) and six items with Likert scale (ranging from strongly disagree (1) to strongly agree (5)), which evaluate quality aspects of the model, such as correctness, consistency, understandability, unambiguousness, completeness, authenticity, flexibility and usability. Regarding the metrics (related to GQM), the results of the items are interpreted based on the median and frequency of response of each item. However, only the responses of Likert scale items will be presented graphically, since all experts answered "no" to the first five items.

Based on the answers of the experts, the internal consistency of the Likert scale items of the questionnaire was analyzed using Cronbach's Alpha method (Cronbach, 1951). The internal consistency verifies how consistent a set of items in a questionnaire is based on the correlation between them (Cronbach, 1951; Raabe; Bombasar, 2020). Thus, Cronbach's Alpha coefficient indicates the degree to which items measure the same quality factor. The coefficient value was $\alpha = 0.87$, indicating a "good internal consistency", according to Devellis (2016). The calculation result of Cronbach's Alpha Coefficient can be viewed in **Appendix C**. Thus, this result indicates that the experts' answers, through the items 6 to 11 (Likert scale) of the questionnaire, are consistent and accurate, in relation to the evaluation of the quality of the AvaliaJS model and its artifacts. We present below the analysis of the results, divided according to the questions (Q1 to Q8) defined in the GQM method.

Q1: Is the AvaliaJS model correct?

Regarding correctness, all experts answered that they did not find errors in the AvaliaJS model (diagrams), canvas model, and assessment project document (item 1). This corresponds to a frequency of 6 positive responses.

Therefore, the experts' answers indicate that the AvaliaJS model is correct.

Q2: Is the AvaliaJS model consistent?

Regarding consistency, all experts answered that they did not find inconsistencies in the AvaliaJS model (diagrams), canvas model, and assessment project document (item 2). This corresponds to a frequency of 6 positive responses. Therefore, the experts' answers indicate that the AvaliaJS model is consistent.

Q3: Is the AvaliaJS model understandable?

Regarding understandability, all experts answered that they did not find confusion in the AvaliaJS model (diagrams), the canvas model, and the assessment project document (item 3). This corresponds to a frequency of 6 responses. Therefore, the experts' answers indicate that the AvaliaJS model is understandable. However, one of the experts commented on the aesthetics of the canvas model, suggesting a better contrast in the background and font colors. This suggestion was not accepted in the current version of the AvaliaJS model but can be analyzed and considered in a future version.

Q4: Is the AvaliaJS model unambiguous?

All experts answered that they did not find ambiguities in the AvaliaJS model (diagrams), in the canvas model and the assessment project document (item 4). This corresponds to a frequency of 6 positive responses. Thus, the experts' answers indicate that the AvaliaJS model is not ambiguous.

Q5: Is the AvaliaJS model complete?

Regarding completeness, all experts answered that they did not miss anything in the AvaliaJS model (diagrams), the canvas model, and the assessment project document (item 5). This corresponds to a frequency of 6 positive responses. Therefore, the experts' answers indicate that the AvaliaJS model is complete. However, one expert pointed out that the understanding of the model and the way how to use it can be improved through a more decomposed description of the technical report (Oliveira; Rocha, 2020a). In this case, the specialist referred to the usability characteristic, in terms of ease of use. However, he/she commented in the field destined to the completeness of the model, as he/she may have missed a more detailed description regarding the use of the model. In this case, the result of the usability evaluation is analyzed in question Q8.

Q6: Is the AvaliaJS model authentic?

When analyzing the authenticity of the AvaliaJS model, based on **Figure 8**, we can observe that most experts strongly agree and agree that the AvaliaJS model diagrams (100%), the canvas model (73.3%), and the assessment project document (100%) adequately include what is necessary to plan the assessment of student performance in a serious game (item 6), as shown in **Figure 8**.

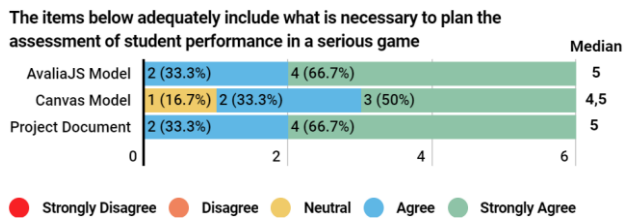

Figure 8. Experts' answers regarding the adequacy (item 6) of the AvaliaJS model.

Nevertheless, some comments on the adequacy of AvaliaJS were made by the experts: (i) use other examples of experimental design in behavior analysis (not only the pre- and post-test classic with an intervention); (ii) the canvas template is not so easy to read/follow. This should probably refer to the completed model (see **Figure 4**). In this case, the template will be analyzed for future updates.

Still on the authenticity of the AvaliaJS model, as shown in **Figure 9,** we can observe that most experts agree and strongly agree that the diagrams of the AvaliaJS model (66.7%), the canvas model (66.7%) and the assessment project document (66.7%) provide more support than the other artifacts that experts know (item 7). Two experts did not evaluate this assertion. As a suggestion, one of the experts comments on adding to the canvas model the blocks "cost", "how well assessed is" and "what will be done with the results" of the assessment. This suggestion should be added in a future version of the AvaliaJS model. Despite the comments, generally, the analyzed data indicates that the AvaliaJS model looks original and can provide support to developers in planning the assessment in SGs.

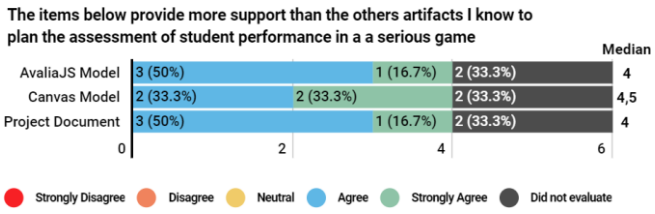

Figure 9. Experts' answers regarding the support (item 7) of the AvaliaJS model.

Q7: Is the AvaliaJS model flexible?

Regarding the flexibility, it was identified that most experts agree and strongly agree that the AvaliaJS model diagrams (100%), canvas model (83.3%) and project document (100%) can be easily adapted to plan the assessment in different learning contexts (item 8), as shown in **Figure 10**.

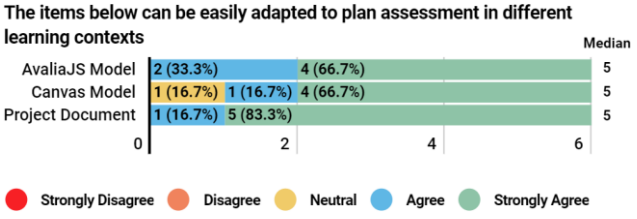

Figure 10. Experts' answers regarding the flexibility (item 8) of the AvaliaJS model.

The results indicate that the AvaliaJS model can be shaped according to the different focus of the game (teaching-learning, training, assessment, motivation and engagement). One expert commented on the possibility of using AvaliaJS in the context of distance learning, which can be analyzed and considered in future studies since the scope of this work is limited to the use of the model in the context of SGs.

Q8: Does the AvaliaJS model have usability?

The usability analysis of the diagrams, canvas model and assessment project document is performed based on the experts' answers in three items: in terms of efficiency (item 9), ease of use (easy to learn to use - item 10) and utility of AvaliaJS (item 11). Regarding efficiency (efficiency is understood as the ability to produce the maximum result with the least effort), according to **Figure 11**, it was observed that 66.6% of the experts strongly agree and agree that the diagrams of the AvaliaJS model allow for assessment planning with minimal effort; 50% agree and strongly agree that the canvas template and the project document allow you to plan the assessment efficiently. The experts scored some observations regarding this item: (1) the way the model is understood and how to use it can be improved through a more decomposed description of the technical report. Although not referring to the efficiency of the model, this comment indicates that a difficult-to-use model indirectly affects the efficiency of assessment planning; (2) one expert comments on the complexity of the evaluation/assessment and questions what would be "minimal effort", which made him strongly disagree with this item. As previously mentioned, the experts are aligned with the literature, about the complexity of the evaluation/assessment. The model proposed was thought to be an aid tool, which can be used to make the assessment planning an activity that can be carried out with more design and documentation and with fewer errors, although it is complex. In general, the model must be applied in the construction of new games so that its efficiency can be measured.

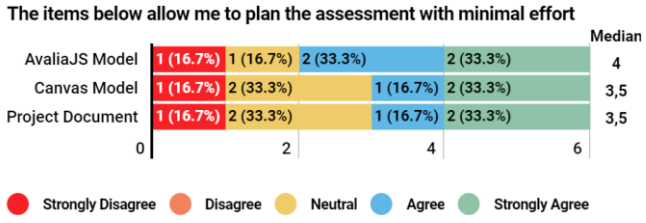

Figure 11. Experts' answers regarding the efficiency (item 9) of the AvaliaJS model.

Regarding the ease of use, which refers to the degree to which the AvaliaJS model is easy to learn, 33.3% of the experts were neutral that the model diagrams are easy to use; 50% were neutral about the ease of learning of the canvas model and assessment project document. The results presented in **Figure 12** may indicate that, because they did not use the AvaliaJS model in practice, the experts were neutral.

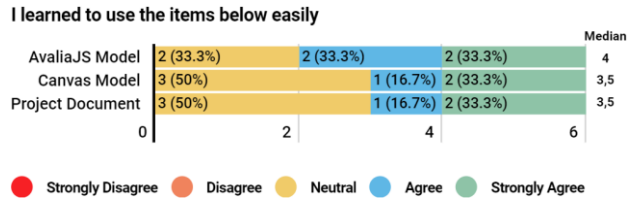

Figure 12. Experts' answers regarding the ease of use (item 10) of the AvaliaJS model.

Some experts pointed out on the ease of use of the model: (1) despite the potential of the three artifacts to guide the assessment planning, they need to be validated in the assessment planning of other new games; (2) need for an interview to clarify possible doubts regarding the use of the model; (3) the colors and subtitles in the AvaliaJS template makes it easier to understand for users. This indicates that the model presents a color division that assists in the identification of contents, since each color corresponds to an aspect of the assessment to be planned, such as design and execution.

Regarding the usefulness of the AvaliaJS model, most experts strongly agree that diagrams (83.3%), canvas model (50%) and assessment project document (66.7%) are useful for assessment planning in SGs, as shown in **Figure 13**. In this context, one of the experts emphasizes the assessment project document as a facilitator in the assessment planning process. This indicates the importance of a model focused on the planning of the design and execution of the student performance assessment since the literature points to a shortage of tools with proposals like this. Furthermore, an expert demonstrated neutrality regarding the usefulness of the canvas model. This can be justified because the expert did not use the model in practice to develop a game assessment plan.

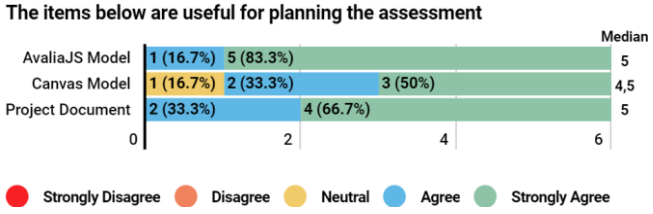

Figure 13. Experts' answers regarding the usefulness (item 11) of the AvaliaJS model.

The results of the quality evaluation by experts indicate that the AvaliaJS model is correct, consistent, clear, unambiguous, complete, authentic and flexible. Regarding usability, the current version of the model needs to be improved to provide excellent usability to those responsible for planning the assessment. The results indicate that the usability of the model needs to be improved in such a way as to be better evaluated. The items of the **third and last part** of the questionnaire sought to know from the experts their considerations about the strengths, weaknesses and additional comments that were identified during the evaluation. As strengths are the description of the content contained in the conceptual model, the intuitive and visual format of the canvas model, the possibility of thinking in advance about the gaps of evaluation and assessment in SGs, among others. The weaknesses are related to the limitations of usability, already scored in this section.

6 Discussions

The conceptual model and its application are discussed through four perspectives, based on Aslan & Balci (2015): (1) artifacts created and final product, (2) methods and processes, (3) people and (4) project.

For the *artifacts created and final product*, the AvaliaJS suggests the use of the canvas template and assessment project document, which should be used to document decisions regarding the planning of the design and execution

of assessment in SGs. As presented in the previous section, the results of the evaluation by the experts indicated a good quality of avaliaJS artifacts (diagrams, canvas template and project document). However, the usability of the AvaliaJS model needs to be reviewed and improved. The diagrams give an overview of the assessment planning process in SGs development life cycle and present some techniques, evaluation/assessment and data collection instruments that can be used. The canvas template and the assessment project document are tools that can be edited and adapted according to the needs of each new project.

Regarding the *methods and processes*, the AvaliaJS model describes, at a high level, the process of conducting evaluation/assessment planning, related to the SGs development life cycle. In summary, the process involves the evaluation/assessment planning with the use of artifacts and the development of the SG, in the pre-production and production stage; and involves the implementation of the evaluation/assessment in the post-production stage. However, the model could specify at a low level some activities and tools needed to facilitate the evaluation/assessment planning process, such as planning and execution activities that produce input and output artifacts. The use of the **5W2H** method (Rossato, 1996) and Falchikov questions (Falchikov, 2005) to construct the canvas model was important for the organization of assessment planning in serious games. Through the questions, it was possible to record the elements and activities, in an agile and holistic way, of the concepts that involve the assessment in SGs. In future works, the model should be expanded to include the (1) costs of the design and execution evaluation/assessment (such as the value of the materials, instruments and artifacts used), predicted by Rossato (1996); the (2) assessment reliability issues (evaluation validation techniques and types of reliability), and the (3) results analysis (techniques for presenting the results and reflecting the educational outcomes achieved), predicted by Falchikov (2005).

Regarding the *people*, the AvaliaJS model describes the actors involved in the evaluation/assessment, during the planning, design and execution (evaluation team, development team, player and observer (teacher/monitor/researcher)) and their roles to perform the process activities. However, future versions of AvaliaJS may add news actors and specify their roles (content writer, analyst, instructional designer, professional expert (of a specific area), game designer, among others), especially in the evaluation and development teams.

Regarding the *project*, three existing serious games were used for analysis and exemplification. However, AvaliaJS should be used in the development of a new project, with the different professionals involved in the evaluation/assessment, and not only with the researchers who created the model. The serious games used to exemplify the AvaliaJS model are focused on teaching-learning, training, engagement, motivation, assessment and feedback. Games focused on awareness, recruitment and rehabilitation were not contemplated. Therefore, the model may not support the planning of serious game assessment with this focus. The three SGs used to exemplify the canvas model did not contemplate the same fill content, because they were games with different types and goals. Thus, it was possible to get a general idea of the content that the AvaliaJS model can cover.

6.1 Threats to validity

We discuss the limitations and threats to validity related to this study, presented below, classified into four categories according to Wohlin et al. (2000).

Internal Validity: The term Serious Games is comprehensive as it involves different purposes and domains. Thus, we believe this could be a threat to the scope of the model. To minimize possible problems, we limited the scope of the model to serious games with a focus on education and training/simulation. Furthermore, the adherence of the model to different theories, techniques, instruments and artifacts for the assessment and evaluation in SGs was verified only in the three games used as proof of concept. Regarding evaluation with experts, we believe that they could be affected by the deadline for evaluating the model. To mitigate this, the experts received the AvaliaJS documents, protocol and questionnaire by email and were able to assess at the most convenient time. Another internal validation problem could be the number of documents that experts had to evaluate. To minimize this problem, we created a technical report (Oliveira; Rocha, 2020) and a protocol with well-defined guidelines and tasks.

External Validity: When selecting people to exemplify the model, there would be a threat to validity if they had not participated in the production of the games analyzed and used as proof of concept, as they might not understand these games. This was mitigated because the researchers were involved in the creation of the analyzed games. This threat could also be minimized if other people were selected to exemplify the model using the games they created. However, the researchers chose to exemplify, because of the difficulty of finding developers who created serious games with different proposals (training, motivation, engagement, learning, evaluation), who assessed the user's performance in these games and who were willing to learn and use the AvaliaJS model. Thus, it was also decided to evaluate by an expert panel, to reduce the threat of bias in the involvement of the researchers in the use of the created model. For this, experts were selected who were also involved in the development of the analyzed games but did not participate in the development of the AvaliaJS. The selected experts evaluated the model and the exemplification but did not use it for planning the assessment in a game, which makes it difficult to measure usability. Another threat to the possibility of generalizing the results is related to the number and experience of specialists. Regarding the sample size (N=6), it is a size that can be used to evaluate and support the development of models or concepts (Beecham et al., 2005). Furthermore, all reviewers are experts and involved in the production of the games used. This made it easier for them to evaluate the fidelity of the artifacts. Therefore, we cannot generalize this assessment to all contexts, for example, to non-specialists.

Construct Validity: The threat regarding the interpretation of model filling (as proof of concept) was mitigated by the involvement of researchers in the production of the games. Another threat is related to the interpretation of the questionnaire by experts. To minimize this problem, the questionnaire used in the expert panel was designed

considering the construction processes of assessment measures and validated through semantic and syntactic analysis with two experts in games and information technology in education. Furthermore, the reliability of the questionnaire was analyzed using the Cronbach coefficient (α=0.87), which indicated good internal consistency.

Conclusion Validity: We do not evaluate the model with the creation of a new game (by external users involved in the production of that game). Thus, the results cannot be considered conclusive, but an indication that the model is functional and that it covers the contents that can be used for evaluation in serious games within the scope of this study.

7 Conclusion

Serious Games are developed for the primary purpose to assess the learning progress and outcomes. However, many of these games are not developed with effective assessment and feedback. This issue could be because of lack of development time, lack of involvement of domain experts in the development process, or focusing on content conveying rather than assessment. In this context, this paper presented and evaluated the AvaliaJS, a conceptual model that supports the planning the design and execution of student performance assessments in SGs. The AvaliaJS has a canvas template, for high-level planning, and a low-level assessment project document, for more specification. Three ready-made serious games were used as proof of concept to analyze and exemplify the use of the conceptual model. The quality of AvaliaJS was confirmed by a panel of six experts, using an online questionnaire for data collection. The internal reliability of the questionnaire was measured using Cronbach's Alpha coefficient α=0.87, which indicates a good consistency and accuracy about the quality evaluation of the AvaliaJS model and its artifacts.

As the main contributions of this study, the conceptual model aims to support the team in the planning, documentation and development of artifacts and data collection in SGs, as well as, in the execution of assessment, learning measurement and constant and personalized feedback for students. The use of the **5W2H** method (Rossato, 1996) and the questions of Falchikov (2005) enabled the identification and organization of assessment elements. The questions that have not been considered can be aggregated into a future proposal for a broader model. The compilation and organization of the theoretical background is also a contribution to the different professionals involved. This is justified by the need to understand this area and to create games that contemplate assessment/evaluation and feedback (technical aspects, such as collections and records, as well as pedagogical aspects, such as theories, techniques and instruments). This also contributes to the "people" pillar (perspective described in Aslan & Balci (2015)) by being better able to design and produce the games.

As future work, the model should be validated in the creation of a game. AvaliaJS can be enhanced if integrated into well-defined processes in each game development phase (planning, analysis, design, implementation, integration and testing, execution and evaluation) and the specification of the roles and actors involved, activities and input and output artifacts. Thus, also in future studies, it is intended to propose a methodology that integrates the conceptual model and artifacts developed in a systematic and facilitating way. In addition, to proposing a computational tool, to support the methodology, to include, customize and analyze the assessment and generate reports of the collected data.

Acknowledgments

This work was carried out with financial support from PROEC/UFABC and CAPES (Process 88887.361026/2019-00).

References

Abdulmajed, H.; Park, Y. S.; Tekian, A. (2015). Assessment of educational games for health professions: A systematic review of trends and outcomes. *Journal Medical Teacher*, v. 37(1), p. 27-32.

Abt, C. C. (1987). Serious Games. Lanham. MD: University Press of America.

Akilli, G.K.; Cagiltay, K. (2006). An Instructional Design/Development Model for the Creation of Game-Like Learning Environments: the FIDGE model. Pivec, M. (Ed.), *Affective and Emotional Aspects of Human-Computer Interaction*: Game-Based and Innovative Learning. IOS Press, p. 93-112.

Alfarah, Z.; Schünemann, H. J.; Akl, E. A. (2010). Educational games in geriatric medicine education: a systematic review. *BMC Geriatrics*, p. 1-5.

Alvarez, J.; Djaouti, D. (2012). Introduction au serious game. 2 ed. Paris: Questions théoriques.

Ariffin, M. M.; Oxley, A.; Sulaiman, S. (2014). Evaluating Game-based Learning Effectiveness in Higher Education. *Procedia - Social and Behavioral Sciences*, v. 123, p. 20-27.

Aslan, S.; Balci, O. (2015). GAMED: digital educational game development methodology. *Simulation*, v. 9(4), p. 307-319.

Baehr, M. (2010). 4.1.2 Distinction between assessment and evaluation. *Faculty guidebook* (p. 7-10). (4th ed.). Lisle, IL: Pacific Crest.

Basili, V.; Caldiera, G.; Rombach, H. (1994). Goal Question Metric Paradigm. *Encyclopedia of Software Engineering*, v. 1, John Wiley & Sons, 1994. p. 528-532.

Battistella, P. E.; Wangenheim, C. G. V.; Fernandes, J. M. (2014). Como jogos educacionais são desenvolvidos? Uma revisão sistemática da literatura. In *XXII Workshop sobre Educação em Computação*, Brasília, DF, p. 1-10.

Beecham, S.; Hall, T.; Britton, C.; Cottee, M.; Rainer, A. (2005). Using an Expert Panel to Validate a Requirements Process Improvement Model. *The Journal of Systems and Software*, v. 76.

Bellotti, F.; Kapralos, B.; Lee, K.; Moreno-ger, P.; Berta, R. (2013). Assessment in and of serious games: an overview. *Advances in Human-Computer Interaction*, v. 2013, p. 1–11.

Bente, G.; Breuer, J. (2009). Making the implicit explicit: embedded measurement in serious games. Ritterfield, U. et al. (eds.), *Serious Games: Mechanisms and Effects*, Routledge, NY, p. 322–343.

Bloom, B. S. (1956). Taxonomy of Educational Objectives: the classification of educational goals – Handbook I: Cognitive Domain. New York, NY, USA: McKay.

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

Calderón, A.; Ruiz, M. (2015). A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education*, v. 87, p. 396–422.

Carey, R. (2015). Game Design Canvas. In: Serious Play Conference, Pittsburgh, PA.

Chaudy, Y; Connolly, T. (2019). Specification and evaluation of an assessment engine for educational games: Integrating learning analytics and providing an assessment authoring tool. *Entertainment Comput.*, v. 30, p. 1-16.

Coelho, J. A. P. M.; Souza, G. H. S.; Albuquerque, J. (2020). Desenvolvimento de questionários e aplicação na pesquisa em Informática na Educação. Jaques, P. A. et al., (Org.) *Metodologia de Pesquisa Científica em Informática na Educação*: Abordagem Quantitativa. Porto Alegre, RS: SBC.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, v. 16(3), p. 297–334.

Csikszentmihalyi, M. (1990). Flow: the psychology of optimal experience. NY: Harper & Row.

Daoudi, I.; Tranvouez, E.; Chebil, R.; Espinasse, B.; Chaari, W. (2017). Learners' Assessment and Evaluation in Serious Games: Approaches and Techniques Review. *International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries*, Xanthi, Greece.

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

Devellis, R. F. (2016). Scale development: theory and applications. 4th ed. Thousand Oaks: SAGE Publications.

Djaouti, D.; Alvarez, J.; Jessel, J. P. (2011). Classifying serious games: the G/P/S model. Felicia, P. (ed.), *Handbook of research on improving learning and motivation through educational games*: Multidisciplinary Approaches, p. 118-136.

Durand, T. (2000). L'alchimie de la compétence. Revue Française de Gestion, n. 127, jan./fev. 2000. p. 84-102.

Emmerich, K.; Bockholt, M. (2016). Serious Games Evaluation: Processes, Models, and Concepts. Dörner R., et al. (eds.), *Entertainment Computing and Serious Games. Lecture Notes in Computer Science*, Springer, Cham, v. 9970, p. 265-283.

Emmerich, K.; Bogacheva, N.; Bockholt, M.; Wendel, V. (2016). Operationalization and Measurement of Evaluation Constructs. In: Dörner R. et al. (eds.), *Entertainment Computing and Serious Games*. Lecture Notes in Computer Science, Springer, Cham, v. 9970, p. 306-331.

Eseryel, D.; Ifenthaler, D.; Ge, X. (2011). Alternative Assessment Strategies for Complex Problem Solving in Game-Based Learning Environments. Ifenthaler D. et al. (eds) *Multiple Perspectives on Problem Solving and Learning in the Digita*l Age. Springer, p. 159-178.

Falchikov, N. (2005). Improving Assessment Through Student Involvement: Practical Solutions For Aiding Learning In Higher And Further Education. New York: RoutledgeFalmer.

Hattie, J.; Timperley, H. (2007). The power of feedback. *Review of Educational Research*, v. 77(1), p. 81-112.

Hays, R. T. (2005). The Effectiveness of Instructional Games: a literature review and discussion. Naval Air Warfare Center Training Systems Division (Technical Report 2005 – 004).

Hettiarachchi, Enosha; Huertas, M. Antonia; Mor, Enric (2013). Skill and Knowledge E-Assessment: A Review of the State of the Art. In IN3 of the UOC Working Paper Series, p. 1-46

Ibrahim, R.; Jaafar, A. (2009). Educational games (EG) design framework: Combination of game design, pedagogy and content modeling. *International Conference on Electrical Engineering and Informatics*, IEEE, Selangor, Malaysia, p. 293-298.

IEEE. (2010). Systems and Software Engineering – Vocabulary. International Standard, ISO/IEC/IEEE 24765:2010(E), p. 1-418.

Ifenthaler, D.; Eseryel, D.; Ge, X. (2012). Assessment for Game-Based Learning. Springer, Verlag New York, p. 1-464.

Jappur, R. F.; Forcellini, F. A.; Spanhol, F. J. (2014). Modelo conceitual para jogos educativos digitais. *AtoZ*: novas práticas em informação e conhecimento, v. 3(2), p. 116-127.

Keller, J. M. (1987). Development and Use of the ARCS Model of Instructional Design. *Journal of Instructional Development*, v. 10(3), p. 2-10.

Keller, J. M. (2009). Motivational Design for Learning and Performance: The ARCS model approach. Springer.

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *Internet and Higher Education*, v. 8, p. 13-24.

Kirkpatrick, D. L.; Kirkpatrick, J. D. (2006). Evaluating Training Programs: The Four Levels. San Francisco, CA: Berrett-Koehler, 3th ed.

Kirkpatrick, J. D.; Kirkpatrick, W. K. (2016). Kirkpatrick's Four Levels of Training Evaluation. ATD Press.

Kolb, A.Y.; Kolb, D.A. (2005). Learning Styles and Learning Spaces: Enhancing Experiential Learning in Higher Education. *Academy of Management Learning and Education*, v. 4(2), p. 193-212.

Korhonen, T.; Halonen, R.; Ravelin, T.; Kemppainem, J.; Kokela, K. (2017). A Multidisciplinary Approach to Serious Game Development in the Health Sector. In *11th Mediterranean Conference on Information Systems (MCIS)*, Genoa, Italy, p. 1-14.

Lopes, M. C.; Fialho, F. A. P.; Cunha, C. J. C. A.; Niveiros, S. I. (2013). Business Games for Leadership Development. *Simulation & Gaming*, v. 44(4), p. 523–543.

Mourão, L.; Meneses, P. P. M. (2012). Construção de Medidas em TD&E. Abbad, G. S.; et al. *Medidas de Avaliação em Treinamento, Desenvolvimento e Educação*: ferramentas para gestão de pessoas. Porto Alegre, RS: Artmed, p. 50-63.

Oliveira, R. N. R. (2020). Planejamento do design e da execução da avaliação do desempenho de alunos em Jogos Sérios. 2020. Dissertation (Computer Science) - Federal University of ABC, Santo André, SP. p. 1-128.

Oliveira, R. N. R.; Belarmino, G.; Rodriguez, C.; Goya, D.; Venero, M. F.; Oliveira Junior, A.; Rocha, R. V. (2019). Avaliações em Jogos Educacionais: instrumentos de avaliação da reação, aprendizagem e comparação de jogos. In *XXX Simpósio Brasileiro de Informática na Educação)*, p. 972-981.

Oliveira, R. N. R.; Cardoso, R. P.; Braga, J. C. B.; Rocha, R. V. (2018). Frameworks para Desenvolvimento de Jogos Educacionais: uma revisão e comparação de pesquisas recentes. In *Anais XXIX Simpósio Brasileiro de Informática na Educação*, p. 854-863.

Oliveira, R. N. R.; Rocha, R.V. (2019). Guerra em Alto Mar: um Jogo de Tabuleiro com Quiz Personalizável para Engajar e Motivar Estudantes. In *XVIII Brazilian Symposium on Computer Games and Digital Entertainment*, p. 1-4.

Oliveira, R. N. R.; Rocha, R. V. (2020a). AvaliaJS: Modelo Conceitual de Planejamento da Avaliação do Desempenho de Alunos em Jogos Sérios. *Technical Report (in Portuguese)*. Federal University of ABC. Available at: https://bit.ly/RelatorioJS. Accessed: May 2020.

Oliveira, R. N. R.; Rocha, R. V. (2020b). Modelo Conceitual de Planejamento da Avaliação do Desempenho de Alunos em Jogos Sérios. In *XVIV Brazilian Symposium on Computer Games and Digital Entertainment*, p. 1-10.

Osterwalder, A. (2004). The Business Model Ontology – A Proposition in a Design Science Approach. Thesis (Business Information Systems) - University of Lausanne, Switzerland, p. 1-172.

Pereira Junior, H. A.; Menezes, C. S. (2015). Modelo para um Framework Computacional para Avaliação Formativa da aprendizagem em jogos digitais. In *XIV Simpósio Brasileiro de Games e Entretenimento Digital*, p. 819-828.

Petri, G.; Wangenheim, C. G. V. (2016). How to Evaluate Educational Games: a Systematic Literature Review. *Journal of Universal Computer Science*, v. 22(7), p. 992-1021.

Petri, G.; Wangenheim, C. G. V. (2017). How games for computing education are evaluated? A systematic literature review. *Computers & Education*, p. 68-90.

Petri, G. (2018). A method for the evaluation of the quality of games for computing education. Dissertation (Computer Science) - Federal University of Santa Catarina. Florianópolis, SC, p. 1-335.

Raabe, A. L. A.; Bombasar, J. R. (2020). Mensuração e testes em Informática na Educação. JAQUES, P. A. et al. (Org.) *Metodologia de Pesquisa Científica em Informática na Educação*: Abordagem Quantitativa. Porto Alegre, RS: SBC.

Rocha, R. V.; Bittencourt, I. I.; Isotani, S. (2015). Avaliação de Jogos Sérios: questionário para autoavaliação e avaliação da reação do aprendiz. In *XIV Simpósio Brasileiro de Jogos e Entretenimento Digital*, p. 1-10.

Rocha, R. V.; Valle, P. H. D.; Maldonado, J. C.; Bittencourt, I. I.; Isotani, S. (2017). AIMED: agile, integrative and open method for open educational resources development. In *XVII IEEE International Conference on Advanced Learning Technologies*, p. 1-6.

Rocha, R. V. (2014). Metodologia iterativa e modelos integradores para desenvolvimento de jogos sérios de treinamento e avaliação de desempenho humano. 2014. Thesis (Computer Science) – Federal University of São Carlos. São Carlos. p. 1 -237.

Rossato, I. F. (1996). Uma metodologia para a análise e solução de problema. Thesis (Production and System Engineering) - Federal University of Santa Catarina. Florianópolis.

Salas, E.; Rosen, M. A.; Held, J. D.; Weissmuller, J. J. (2009). Performance measurement in simulation-based training: a review and best practices. *Simulation & Gaming*, v. 40(3), p. 328–376.

Salen, K.; Zimmerman, R. (2003). Rules of Play: Game Design Fundamentals. MIT Press.

Sarinho, V. T. (2017). Uma Proposta de Game Design Canvas Unificado. In *Anais XVI Simpósio Brasileiro de Jogos e Entretenimento Digital*, p. 141-148.

Savi, R.; Gresse Von Wangenheim, C.; Ulbricht, V. R.; Vanzin, T. (2010). Proposta de um modelo para avaliação de jogos educacionais. *RENOTE*, v. 8(3), p. 1-10.

Shell, J. (2008). The Art of Game Design: A Book of Lenses. 2nd ed. AK Peters.

Slussareff, M.; Braad, E.; Wilkinson, P.; Strååt, B. (2016). Games for Learning. DÖRNER, R. et al. (Eds.), *Entertainment Computing and Serious Games*, LNCS, v. 9970, p. 189–211.

Sousa, T. C. (2014). Manual do Game Design Canvas. Available at: http://bit.ly/gdcanvas. Accessed: Jan 2019.

Star, K., Vuillier, L.; Deterding, S. (2016). D2. 6 Prosocial Game design methodology. *Gamification of Prosocial Learning for Increased Youth Inclusion and Academic Achievement*, p. 1-60.

Trybus, J. (2010). Game-Based Learning: What it is, Why it Works, and Where it's Going. NMI White Paper. New Media Institute, New York.

Victal, E. R. N; Menezes, C. S. (2015). Avaliação para Aprendizagem baseada em Jogos: Proposta de um Framework. In *Anais XIV Simpósio Brasileiro de Jogos e Entretenimento Digital*, p. 970-977.

Walker, E. (2015). Game Canvas Design. Ideas at Play Workshop 2015. Available at: drive.google.com/file/d/0B66VbdUcz8k9TVotdEFJcFZQeE0/view. Accessed: May 2019.

Wang, R.; De Maria, S. J.; Goldberg, A.; Katz, D. (2016). A Systematic Review of Serious Games in Training Health Care Professionals. *Simulation in Healthcare*: The Journal of the Society for Simulation in Healthcare, v. 11(1), p. 41-51.

Wangenheim, C. G. V.; Kochanski, D.; Savi, R. (2009). Revisão Sistemática sobre Avaliação de Jogos Voltados para Aprendizagem de Engenharia de Software no Brasil. In *Anais Fórum de Educação em Engenharia de Software*, Fortaleza, CE, p. 1-8.

Westera, W.; Nadolski, R.; Hummel, H.; Wopereis, I. (2008). Serious games for higher education: a framework for reducing design complexity. *Journal of Computer Assisted Learning*, v. 24, p. 420–432.

Wohlin, C.; Runeson, P.; Höst, M.; Ohlsson, MC.; Regnell, B.; Wesslén, A.. (2000). Experimentation in Software Engineering-An Introduction. (1st ed.). Kluwer Academic Publishers, Norwell, MA, USA.

Yedri, O. B.; Aachak, E. A.; Bouhorma, M. (2018). Assessment-driven Learning through Serious Games: Guidance and Effective Outcomes. *International Journal of Electrical and Computer Engineering*, v. 8(5), p. 3304-3316.

Zerbini, T.; Borges-ferreira, M. F.; Abbad, G. S. (2012). Medidas de Reação a Cursos a Distância. Abbad, G. S. *Medidas de Avaliação em Treinamento, Desenvolvimento e Educação:* ferramentas para gestão de pessoas. Porto Alegre, RS: Artmed, p. 91-107.

Zinovieff, M. A.; Rotem, A. (2008). Review and analysis of training impact evaluation methods, and proposed measures to support a United Nations system fellowships evaluation framework prepared. *WHO's Department of Human Resources for Health*, p. 1-46.

Zyda, M. (2005) From visual simulation to virtual reality to games. *Computer*, v. 38(9), p. 25–32.

Appendix A - Summary of the analyzed games

Table 3. Summary of the canvases of the analyzed games.

Canvas \ Game	G1	G2	G3
Purposes	Training Assessment	Motivation Engagement	Teaching-Learning Assessment
Planning	Full game	Full game	Game phase
What will be assessed during the game? (focus)	Training Reaction *Feedback*	Teaching-Learning Reaction *Feedback*	Teaching-Learning Reaction *Feedback*
Why should the assessment be carried out? (assessment purposes)	Knowledge Skill Attitude Motivation Engagement Self-assessment Learning Report Game Progress User Interaction	Knowledge Motivation Engagement Learning Report Game Progress	Knowledge Skill Attitude Commitment Motivation Engagement Learning Report Game Progress User Interaction
How should the assessment be carried out? (theories, techniques, instructions and artifacts	- Theories - Bloom's Taxonomy - Pre-test (game phase) - Observation - Theories (reactions) - Reaction and self-assessment (Questionnaire) - Human error dimensions - Theories about feedback - Player profile questionnaire and registration - *Debriefing* - Game phase with data collection of user actions	- Theories - Observation - Theories (reactions) - Reaction (Questionnaire) - Theories about feedback - Player profile questionnaire - Data collection of the user: data collection sheet	- Theories - Pre/post-test - Observation - *Think-Aloud* - Interview - Theories (reactions) - Reaction (Questionnaire) - Human error dimensions - Theories about feedback - Player profile questionnaire - *Debriefing* - Game phase with data collection of user actions
When should each assessment be carried out? (time/ duration/deadline)	Total of 1h30: 30 min for each activity: answer questionnaire, game interaction and debriefing	Total of 3h: 10 min- profile questionnaire; 15 min- video display with rules; 130 min- game interaction; 25 min- reaction (questionnaire)	Total of 1h: 5 min- profile questionnaire; 5 min- pre-test; 30 min- game interaction;10 min - post-test / reaction (questionnaire); 10 min- *debriefing and interview*
Where will each assessment be carried out? (mode, place, context and equipment)	**Mode**: online; **Place**: fire station; **Context**: training session; **Equipment**: PC.	**Mode**: offline; **Place**: laboratory; **Context**: monitoring, workshop, class, course; **Equipment**: analogic game and stopwatch.	**Mode**: online; **Place**: laboratory, residence; **Contexto**: workshop, class/course; **Equipment**: PC.
Who is involved in carrying out the assessments? (participants / role)	**Players**: firefighters (soldier, corporal and sergeant); **Observer**: researcher.	**Players**: students; **Monitor**: mediates and monitors responses; **Observer**: researcher.	**Players**: undergraduate students; **Observer**: researcher.

Appendix B - Questionnaire of quality evaluation of the AvaliaJS Model

Table 4. Questionnaire of quality evaluation of the AvaliaJS

Part I- Expert's Profile	
Items	**Scale**
1. How many serious games (digital or non-digital) have you developed?	Multiple Choice: () None; () Less than 5 serious games; () 5 to 10 serious games; () More than 10 serious games.
2. How many serious games (digital or non-digital) have you used?	
3. How many serious games (digital or non-digital) have you evaluated?	
4. How do you usually evaluate serious games?	() Competencies Assessment (knowledge, skill, attitude); () Commitment Assessment; () Motivation Assessment; () Engagement Assessment; () Self-Assessment; () Pre/Post-test; () Observation; () Interview /Debriefing; () I don't usually evaluate; () Others
5. Do you know any artifact of planning the design and execution of the student performance assessment in serious games?	Open-ended
6. Comment on the degree of difficulty in planning the assessment of reaction and learning, inside and outside serious games	
Part II - Items related to quality characteristics of AvaliaJS (diagrams, canvas template and project document)	
a. **Correctness**: Degree of how correct the model is, what is the extent of existing errors.	
1. Did you find any errors in AvaliaJS?	() Yes or () No
b. **Consistency**: Degree of uniformity, standardization, and freedom from contradiction among the components of AvaliaJS	
2. Did you find any inconsistencies in AvaliaJS?	() Yes or () No
c. **Understandability**: Degree to which the purpose, concepts, and structure of the AvaliaJS are clear to the experts.	
3. Did you find anything confusing in AvaliaJS?	() Yes or () No
d. **Unambiguousness**: Degree to which a definition/statement is described in terms that only allow a single interpretation.	
4. Did you find any ambiguity in AvaliaJS?	() Yes or () No
e. **Completeness**: Degree of coverage of the AvaliaJS, if the model is sufficiently complete.	
5. Did you notice something missing in AvaliaJS?	() Yes or () No
f. **Authenticity**: Degree to which the AvaliaJs can realistically represent the domain it was defined.	
6. The AvaliaJS adequately includes what is necessary to plan the assessment of student performance in a serious game.	Likert Scale: 1 (Strongly Disagree) to 5 (Strongly Agree)
7. The AvaliaJS provides more support than the other artifacts I know for planning the assessment of student performance in a serious game. Do not respond if you do not know other artifacts.	
g. **Flexibility**: Degree to which the AvaliaJS can be adapted to changes, allowing it to be applied in contexts other than the one defined.	
8. The AvaliaJS can be easily adapted to plan assessment in different learning contexts.	Likert Scale: 1 (Strongly Disagree) to 5 (Strongly Agree)
h. **Usability**: Degree of understanding, ease of use and applying the AvaliaJS in an effective and efficient way.	
9. The AvaliaJS allows me to plan the assessment with minimal effort.	Likert Scale: 1 (Strongly Disagree) to 5 (Strongly Agree)
10. I learned to use the AvaliaJS easily.	
11. The AvaliaJS are useful for planning the assessment.	
Part III - Final Considerations	
What are the strengths of the AvaliaJS model and its artifacts?	Open-ended
What are the weaknesses of the AvaliaJS model and its artifacts?	
Do you have any more comments, suggestions and criticisms about the AvaliaJS?	

Appendix C - Calculation of Cronbach's Alpha Coefficient

Table 5. Calculation of Cronbach's Alpha Coefficient

E	Item 6			Item 7			Item 8			Item 9			Item 10			Item 11			Total
	D	CM	PD	D	CM	PD	D	CM	PD	D	CM	PD	D	CM	PD	D	CM	PD	
1	5	5	5	4	5	4	5	5	5	3	3	3	3	3	3	4	4	4	**73**
2	5	5	5	0	0	0	5	5	5	4	3	4	4	4	4	5	5	5	**68**
3	5	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	**85**
4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	**90**
5	4	4	5	0	0	0	4	3	5	4	4	3	4	3	3	5	4	5	**60**
6	4	3	4	4	4	4	4	4	4	1	1	1	3	3	3	5	3	4	**59**
Var	**0.3**	**0.7**	**0.3**	**5.0**	**5.6**	**5.0**	**0.3**	**0.7**	**0.2**	**2.3**	**2.3**	**2.3**	**1.0**	**3.0**	**1.0**	**0.2**	**0.7**	**0.3**	**164.3**

Caption: **D** - Diagrams | **CM** - Canvas Model | **PD** - Project Document | **E** - Expert | **Var** - Variance

Number of items (k)*	**18**
Sum of item variances ($\sum V_i$)	**28,6**
Total test variance (V_T)	**164,3**
Cronbach's Alpha (α)	**0,87**

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum V_i}{V_T}\right)$$

Alpha Value	Internal consistency (Scales)
α >= 0,9	Excellent
0,9 > α >= 0,8	Good
0,8 > α >= 0,7	Acceptable
0,7 > α >= 0,6	Questionable
0,6 > α >= 0,5	Poor
α <= 0,5	Unacceptable

* Each item of the questionnaire evaluated three artifacts (diagrams, canvas model and project document). Therefore the total number of items is equal to 18.