Article

A New Measure for Serious Games Evaluation: Gaming Educational Balanced (GEB) Model

Kim Martinez [1] , María Isabel Menéndez-Menéndez [1] and Andres Bustillo [2,*]

1 Department of History, Geography and Communication, Universidad de Burgos, 09001 Burgos, Spain
2 Department of Computer Engineering, Universidad de Burgos, 09006 Burgos, Spain
* Correspondence: abustillo@ubu.es

Featured Application: A metric to evaluate gaming and educational features of already developed serious games. This model can also guide the design of new serious games.

Abstract: Serious games have to meet certain characteristics relating to gameplay and educational content to be effective as educational tools. There are some models that evaluate these aspects, but they usually lack a good balance between both ludic and learning requirements, and provide no guide for the design of new games. This study develops the Gaming Educational Balanced (GEB) Model which addresses these two limitations. GEB is based on the Mechanics, Dynamics and Aesthetics framework and the Four Pillars of Educational Games theory. This model defines a metric to evaluate serious games, which can also be followed to guide their subsequent development. This rubric is tested with three indie serious games developed using different genres to raise awareness of mental illnesses. This evaluation revealed two main issues: the three games returned good results for gameplay, but the application of educational content was deficient, due in all likelihood to the lack of expert educators participating in their development. A statistical and machine learning validation of the results is also performed to ensure that the GEB metric features are clearly explained and the players are able to evaluate them correctly. These results underline the usefulness of the new metric tool for identifying game design strengths and weaknesses. Future works will apply this metric to more serious games to further test its effectiveness and to guide the design of new serious games.

Keywords: serious games; game evaluation; game design; metrics

1. Introduction

Learning is one of the most important functions for the development of society. Constant innovation and adaptation of instructive tools to new trends are necessary for most important educational topics, in both primary and university education. One educational topic concerns video games, which play a relevant role in society as an industry and as a hobby enjoyed by a large part of the population. Reaching many target audiences through this medium has proven to be very productive in different fields [1]. Players commit themselves to a game when they are enjoying the experience and therefore engage with its educational content (engagement) [2]. Furthermore, assimilating this knowledge through narrative and fun gameplay is more effective than traditional education [3].

Serious games have been developed on a range of topics within surprisingly varied fields, from indie games to academic works [4]. Successful achievement of the instructional objectives proposed in many video games has been demonstrated [3]. However, introducing a game to the target audience is not enough to achieve the learning aims. If it is not a fun experience, then the hedonic impulses of users will not be satisfied, so their attention may be distracted or they may stop playing. If there is no engagement, learning is more difficult for users. Additionally, if the educational content is not well adapted to the story or the gameplay, players will not be able to learn. Likewise, the distribution of these

contents throughout the game contributes to a progressive and an interesting experience. For these reasons, an evaluation system is necessary to assess the instructional capacity of a game. Academic metrics or measures on the design of serious games are essential tools for future research.

A review of recent work [5] shows how there already are some studies that divide the evaluation in learning and gaming aspects. Some work focused on the learning evaluation using the Serious Game Design Assessment (SGDA) framework [6]. Other work assessed the gameplay experience, using the Conceptual Research Model [7], and the Games, Motivation and Learning study [8]. However, these studies usually lack one of the two aspects of assessment, or do not consider their balance [5]. Another main inconvenience is that these models offer only qualitative analysis, not allowing a score for the resulting design. Therefore, the objective of this paper is the development of a tool that researchers and educators may use to quantitative value a developed, serious game.

The Gaming Educational Balanced Model (GEB) is developed from the Mechanics Dynamics Aesthetics (MDA) framework [9] and the Four Pillars of Educational Games (4PEG) theory [10]. MDA is focused on the gameplay elements, but it overlooks a game's educational content, and in addition to that, it is not developed for design evaluation content. The 4PEG theory does evaluate playability and learning aspects in a generalized way. However, this study lacks a detailed evaluation of both aspects and the balance between them. Moreover, it has no easy-to-follow rubric, leaving the scoring to the evaluator's personal decisions and experiences.

The main novelty of the new GEB model is that it overcomes these limitations. The GEB model can be used to evaluate the instructional capacity of the game from all perspectives. Compared with previous models, it assigns similar importance to both fun and learning content; the scores of each section are of equal value, and in addition, their distribution and the balance between fun and learning are also rated. This new model defines all the features a game needs to engage a player and to adapt its instructional capacity accordingly. The result is a metric or measure for immediate use to evaluate serious games as educational tools. In addition, as a result of the description of each feature that forms a serious game, the GEB model can also be used as a guide. Designers can follow the indications of each feature in their high scores to know how to implement each ludic and educational element in their serious game.

As well as proposing this model, the new measure was tested by one of the authors and 36 volunteers, applying it to three indie serious games to evaluate their instructional capacity. The number of games was decided due to the available volunteers, so that each one had 12 completed surveys. These three games were chosen because of their aim to raise social awareness about mental illnesses, since volunteers study degrees dedicated to media. This is a promising field of application, although the production of serious games is still very limited [11]. For this reason, indie serious games were searched for on the Steam platform, following two criteria, and the ones that best fit were chosen. First, each one is based on a different genre (puzzle platform, graphic adventure and survival horror), giving insight into possible contrasts between the gameplay and the educational content. Furthermore, the game experience is limited to a few hours, so its effectiveness could easily be verified in a workshop.

This paper will be structured as follows: In Section 2, the MDA and 4PEG models will be described as the foundation for the development of the measure in the GEB model, besides a definition will set out every GEB model section and an explanation of the adaptation of some features and the removal of others; an explanation of the results of applying the GEB model to the three indie games will be detailed in Section 3; finally, the conclusions and future lines of work will be presented in Section 4.

2. Materials and Methods

Section 2.1 introduces the base frameworks of the GEB model, the 4PEG and the MDA, and the sections into which they are divided. Section 2.2 begins by explaining how these sections have been adapted in the GEB model in order to evaluate all aspects of a serious game with their optimal scoring. Then, Sections 2.2.1–2.2.3 develops the Game Overview, Educational Content and Overall Balance headings of the GEB model, with all the features that compose them. These are mostly derived from the 4PEG and MDA frameworks, which are explained and further developed with the help of the most current and relevant literature.

2.1. Previous Serious Games Studies: MDA and 4PEG Models

To start developing the GEB model, a literature review was carried out on previous studies that had developed quantitative metrics for serious games. The chosen terms were "serious games AND metric OR rubric". Scopus gave 241 results, while Springer showed 397. However, the screening process of these 638 results only resulted in one eligible option: the 4PEG theory [10]. There were some studies [12,13] that compiled important elements of game design, but none offered a quantitative metric to measure them and, therefore, they were not useful for this paper.

The 4PEG theory considers both educational content and gameplay. The author also developed the Magic Bullet concept, that explained the need to find the right balance between the game context and the learning context [14]. Finally, available resources for classroom applications are also taken into account. The theory establishes a set of measures for the different components, resulting in a total score out of 100 that values the instructional capacity of the game.

The model also classified four categories of learning: (1) Things that the game developers deliberately design and that can be learnt; (2) things that must be learnt to achieve the goal of the game that will usually be a subset of the first category; (3) external learning outside of the game; and, (4) coincidental learning, which includes any other knowledge triggered by the game. The purpose of this theory was to perform a structured analysis through its components and subsections that are shown in Figure 1. The results constituted an assessment of both commercial and serious games and their potential as learning instruments within the classroom. The percentages evaluated in the 4PEG theory are explained below.
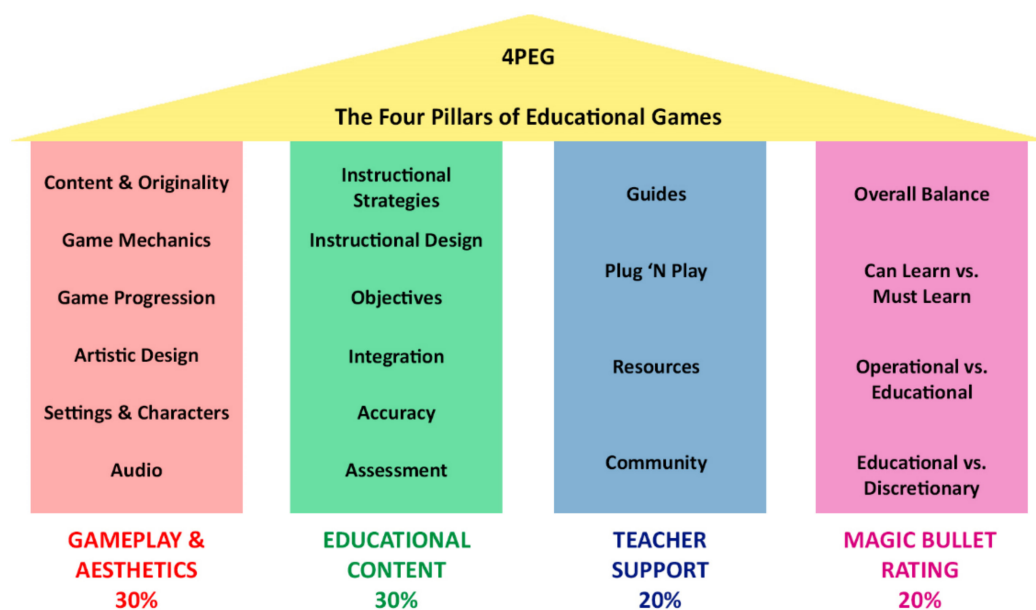

Figure 1. Component of 4PEG Model.

The fun of the game and the gaming experience was rated under the first pillar: Gameplay and Aesthetics. This section represented 30% of the overall rating and was considered important, because, if the player feels no pleasure, engagement with the game will be harder. The total percentage was divided between six aspects, each of equal value. Any surprising content and general adaptiveness to the genre was measured under Content and Originality. Every possible action that the player can perform, its logic and its difficulty were assessed under Game Mechanics. The transition throughout the story and the different levels were rated under Game Progression. The attractiveness of the graphic display and its appeal for the game genre and type of desired player were valued under Artistic design. Settings and Characters refer to the adaptiveness of these components to the total experience. Audio refers to a review of the adjustment of music and sounds to the gameplay and represents an evaluation of its appeal to the player.

These aspects of the gameplay, although not wrong, were not well defined according to the most current research. For this reason, it was decided to replace this section with the MDA framework [9]. Game design tends to refuse a common methodology for the development of serious games. However, a recent study [15] defines MDA as the closest approach to a unified framework. This theory defines the main characteristics that affect the gameplay and its engagement from the perspectives of both the designer and the player. The objective is to elicit emotional responses that assist comprehension of the learning content. The layers that form this model are: (1) mechanics: all possible player actions and their representations; (2) dynamics: predictable runtime behaviors that emerge from mechanics; (3) aesthetics: audiovisual responses that evoke emotions in the player [15]. Other works have been determining which elements are needed in these three layers, such as the Dynamical Model of Gamification of Learning (DMGL) [16]. This work links every MDA feature to the sensations required to attract users. Furthermore, the DPE framework [17] added the technology and User eXperience (UX), essential for player interaction, to the MDA [9]. Although these studies are not dedicated to the evaluation of serious games, they do identify all the engagement features that are necessary.

The three following pillars of 4PEG are part or the educational overview that accounts for the remaining 70%, as shown in Figure 1. The Educational Content pillar that accounts for 30% of the overall score was also divided into another six equally valued features. Instructional Strategies are focused on how well the gameplay is adapted to the intended learning outcomes. Instructional Design covers all the game elements that engage with the proposed problem, activate the knowledge to solve it, and demonstrate, use and integrate the new knowledge. The inclusion of the learning aims in the content was measured under Objectives. The extent to which the desired learning outcomes formed part of the required interactions during the game was rated under Integration. Accuracy was centered on measuring the reliability of both the concepts and the principles that are needed to achieve the goal. Assessment is part of the game system that measures the ability of the player to achieve the objectives.

The third and fourth pillars both account for 20% of the total score that is distributed equally in four sections. The layout of the game in terms of the educational resources available to both professors and students is assessed under Teacher Support. In this section, the existence of Guides with the content and process description are evaluated, as are Plug and Play materials that include lesson plans. Any additional information specifically for teachers was rated under Supplementary Resources. The creation of digital forums for sharing ideas and experiences was rated under Community. Finally, the Magic Bullet Rating reflected an evaluation of the balance between different kinds of learning. Overall Balance was focused on the relationship between the four pillars to achieve the game goal. Can Learn vs. Must Learn is a rating of the inclusion of these contents within the actions in the game. Time spent on the game in relation to the learning purpose was assessed under Operational vs. Educational. Finally, an appropriate balance between learning and fun was assessed under Educational vs. Discretionary.

2.2. Gaming Educational Balanced Model

The GEB model was developed from the 4PEG Model structure. Nevertheless, 4PEG does not have interrelationships and percentages justified by serious games design research. So, the GEB model reviewed and modified the total percentages of each section. These changes respond to a search for recent studies [3,16,18–23] that demonstrate the importance of balancing gameplay and educational content. In addition, the relevance of the elements that compose them was also analyzed to give them the appropriate ratios.

While 4PEG is divided only into Game Overview and Educational Overview, the GEB model adds a third section: Overall Balance. In 4PEG theory, the balance between both sections responds to the Magic Bullet concept [10] and enters into the Educational Overview. However, several studies [21,24,25] have shown the need to add a third perspective. A game can be fun and engaging on the one hand, and properly introduce educational content on the other. Moreover, if these contents are not distributed in time and in the gameplay, motivation is reduced [26] and the player's cognitive load is increased [27]. Therefore, this section is relevant, although it does not have the same importance as the two main ones.

The new measure, called the GEB metric, distributes the percentages in the same measure for Game and Educational Overview. This decision follows the line of several studies [20,25,28,29] that have developed theories about the need for a balance of seriousness and playfulness in game design. Furthermore, this perspective is not only shared by researchers and designers. Players perceived the same needs in their engagement with learning and play [22,30,31]. Therefore, Game Overview and Educational Overview receive 40% of the total percentage each. While Overall Balance keeps 20%, which is half of the previous ones, but still receives 1/5 of the total score for the effect it causes. The changes between the 4PEG and the GEB ratios can be seen in Figure 2.
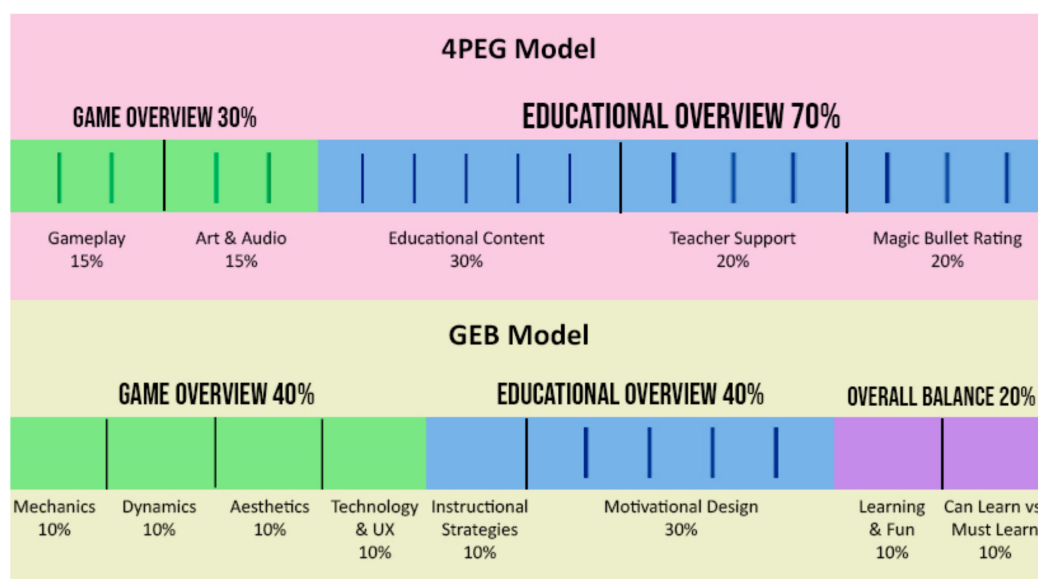

Figure 2. Comparison between 4PEG Model and GEB model.

The distribution of Game Overview has given 10% to its four sections: Mechanics, Dynamics, Aesthetics, and Technology and UX. Both in the definition of the MDA framework made by Hunicke et al. [9] as in its later use by numerous studies [15], the relevance of each layer has always been equivalent. Each feature has many elements that are essential for player engagement and cannot be prioritized over others. In the same way, Technology and UX has been a section that later models [9,17,32] have added as the fourth pillar of gameplay. This feature contemplates the direct perception of the player's motivations and acquires the same importance [33].

Educational Overview gives different percentages to its two sections: 10% of the score to Instructional Strategies and 30% to Motivational Design. In the game design research

on the application of learning, the relevance given to these two sections is distinguished. Although the construction of the learning context is important, the aspects directly related to the motivation created in the player determine the level of learning [34]. Motivated users feel less pressure and emotional exhaustion [35]. They are also more willing to learn and give more value to the experience, paying more attention and getting more positive reactions [36]. Therefore, it is concluded that Motivational Design has three times the score of Instructional Strategies due to its need to not burden the players and help them learn [37]. Within Motivational Design there are five different features and each one receives an equivalent 6% of the score. The game design theories that focus on this educational design [20–23] give the same relevance to all its contents.

Finally, the Overall Balance is split into two subsections quantified both as 10% of the total score: Learning and Fun and Can Learn vs. Must Learn. The equitable distribution responds to the cognitive and emotional needs of the player since the game and educational features must be applied in a stable and flexible way [25]. In this perception, its inclusion in time is as important as the differentiation of knowledge between its mechanics [21].

Once the main sections of the GEB model are obtained, the design characteristics that each feature must have are defined. These are obtained from the same studies that have made it possible to establish the ratios [3,9,16–26,28,29]. The features are listed and the information between studies is compared to obtain their particularities. As the GEB metric is part of the 4PEG, the elements are also compared between them, resulting in them being adapted, added or eliminated for the definition of the metric. When the final features are obtained, Sections 2.2.1–2.2.3 explain the game design needs following the analyzed bibliography. As a result, the GEB metric is obtained to measure serious games, and serve as a guide to design and to develop new serious games using the optimal features that have been described.

The metric and rubric for applying the GEB model to any game are attached as Supplementary File S1, and an overview can be seen in Table 1. The following headings summarize every attribute and its necessary characteristics that fulfil the goals of a serious game. These evaluation characteristics arise from the development of each feature that occurs in Sections 2.2.1–2.2.3. The correspondence with each point is indicated in each one. Likewise, the summary of possible scores that a serious game can have in every feature are shown under the Punctuation heading. The exact assignment of the score of each attribute is explained in Supplementary File S1.

Table 1. Summary table of GEB Model.

Section	Feature	Punctuation	Evaluation
Game Overview	Game Mechanics	10/8/5/2/0	Levels, missions, rules, abilities, scores and leaderboards. Prizes and objects that are received when the objectives are achieved, as well as the responses of the characters, the choices of the players and the audios and videos.
	Game Dynamics	10/8/5/2/0	Progressive difficulty, time patterns, possible strategies and scheduled rewards. Missions or characters to interact with, besides objects that can be unlocked or explored, and the possibilities of competition and cooperation.
	Aesthetics	10/8/5/2/0	Emotions such as love, beauty, friendship, delight, surprise, honor, laughter, envy, drama, chills or surprise.
	Technology and User Experience	10/8/5/2/0	Interface must motivate to play and be pleasant, also offering decision-making, self-perception of the actions and objectives of the game to give a sense of control. Everything in the short and long term.

Table 1. *Cont.*

Section	Feature	Punctuation	Evaluation
Educational Content	Instructional Strategies	10/8/5/2/0	Tests to acquire knowledge that avoid genres associated with learning, and introduce educational content through gameplay and narrative. In addition, the gameplay must be more relevant than the educational part.
	Backstory and Production	6/5/3/1/0	An interesting plot, that does not interrupt the fun to introduce the educational content, and that allows to choose its development.
	Realism	6/5/3/1/0	Realistic graphics, sounds, and dynamics are engaging and help to focus. There are avatar customization options, and character interactions are expressive.
	AI and Adaptivity	6/5/3/1/0	AI or game interactions adapt features and difficulty to the player.
	Interaction	6/5/3/1/0	Interactions are easy and motivating, taught in a tutorial and progressively more complicated. In addition, there is collaborative interaction.
	Feedback and Debriefing	6/5/3/1/0	Feedback comes through rewards and character interactions. The teacher can help during the game, and a debriefing takes place.
Overall Balance	Learning and Fun Balance	10/8/5/2/0	It's fun while educational. Additionally, learning content is distributed throughout the game's narrative while the controls become progressively more difficult.
	Can Learn vs. Must Learn	10/8/5/2/0	Important learning is implemented in the main objectives. There is extra learning that is introduced in the exploration possibilities. There is not a large volume of information to distract from the game.

2.2.1. Game Overview

This section lists the different characteristics associated with the gameplay. Each one affects the capacity of the game to engage players towards achieving the proposed goals. Three of the four subsections are based on the MDA framework [9], that classifies the main components of gameplay. The labels relate to the 4PEG model Game Mechanics and Artistic Design and Audio, presented by Aesthetics, as shown in Figure 3. In addition, there is another subsection on Technology and User Experience. This aspect has been added to the Game Overview, due to the effect experienced by the player in their interaction with the technology [38] and the interface [39,40].
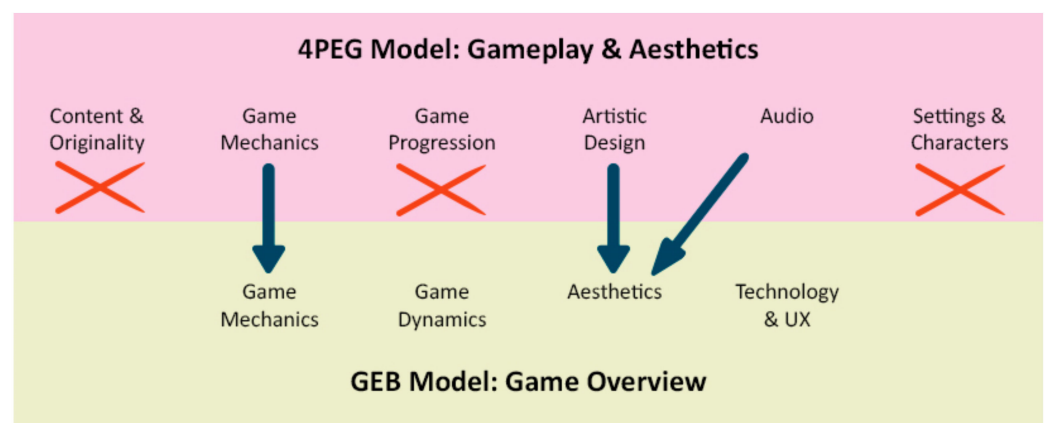

Figure 3. Adaption of 4PEG Model Gameplay and Aesthetics to GEB model Game Overview.

The requirement of the user's intrinsic motivation to play a game must be emphasized in any explanation of the importance of these features [23,36]. Introducing the correct gameplay elements in the experience will turn a serious game into a successful and a fun experience. The GEB model defines the MDA subsections in line with the DMGL theory of Kim and Lee [16]. These authors determined the game components that elicited the emotions under each label to engage the player. According to Malone [41] these emotions or feelings are the key characteristics of a learning game (KCLG): challenge, curiosity and fantasy. Challenge is created by having clear goals that are relevant to the user. Curiosity is promoted using the game mechanics and aesthetics for immersion in the narrative. Fantasy is the environment that evokes mental images that activate emotions and logical thought processes. Malone and Lepper [42] added another characteristic derived from the correct application of the previous ones: The sensation of control, a feeling of greater self-determination that a learner progressively experiences.

Game Mechanics

The first MDA subsection, Game Mechanics, encompasses the game components that manage actions, behaviors and control. Subsequently, dynamics and aesthetic experiences are created for the player [9]. The characteristics of the mechanics of the game, which create the challenges for the players, are related to a sense of difficulty. These factors are levels, objectives and missions, rules, abilities, scores and leader boards [43]. The features of mechanics offer rewards and feedback that build up the fantasy of the story [36]. Medals, prizes, and in-game items that are received when the objectives are achieved, together with character responses, player choices and the audios and videos that are displayed [16].

Game Dynamics

Following the MDA framework, there are Game Dynamics. This layer can be described as the systemic patterns of behavior when using mechanics during the gameplay. In turn, dynamics create aesthetic experiences that the player selects during the game [9]. The characteristics of Game Dynamics are related to the proposed challenges and the responses of these patterns to the player actions [25]. Its components include difficulty progress, time patterns, possible strategies, programmed rewards and relationships with the features of mechanics [16]. The features of dynamics arouse curiosity, pointing as they do to new objectives and possibilities within the game environment [23]. These can be missions or characters with which the player must interact, items to unlock or explore, competition and cooperation possibilities [24,35].

Aesthetics

The last MDA subsection is Aesthetics [44], which includes all the emotional responses evoked in the player by the visual and audio components. From the player's perspective, aesthetics defines the dynamics and the operable mechanics [9]. The features of Aesthetics are responsible for creating a sense of fantasy, due to the arousal of positive and powerful emotions that are associated with love, beauty, friendship, delight, surprise and honor [45,46]. Its features also relate to curiosity, adding either a positive or a negative perspective, by calling attention to the factors that can provoke them in the game. Among such sensations are laughter, envy, drama, shuddering and surprise [47].

Technology and User Experience

The last subsection in the Game Overview is Technology and User Experience. The Technology layer corresponds to the software and hardware tools that enable user interactions with the game. These features determine the development and design of any game from the beginning of the process. Likewise, the UX layer is mainly affected by the Technology layer [9]. The UX component can be defined as the effect on the perceptions of each player during interactions with the game interface [39,40,47]. User Experience is

based on both the graphical interface of the device in use and the experience of player interaction through that interface [48].

A serious game that offers a positive UX must have the pleasure factor carefully balanced with the fun factor, in both the short and the long term [33]. There are four aspects to take into account: motivation, choices, usability and aesthetics [44]. An initial interest in the pleasure of interaction maintains motivation in the short-term and the offer of rewards prolongs it in the long-term [32]. Choices define the game structure and the rules that players have to follow to make decisions. Among others, these can be short-term tactics and long-term strategies, requiring greater skill [49]. The following feature is usability that facilitates the self-perception of actions and objectives in the game. User Experience is designed to give a sense of control over immediacy and the option of mastering the game in the long-term [47]. Finally, the interface aesthetics relates to all the visible elements that the player experiences. These must be pleasant from the beginning and they continue in this way as the game progresses [50].

2.2.2. Educational Overview

The Educational overview section evaluates the inclusion of educational content in gameplay. Its aim is to create interesting, positive and didactic experiences for users defining certain components and rules in the game development. The educational metrics are divided into Instructional Strategies and Motivational Design. Instructional Strategies relates to the adaption of learning content to the gameplay [47] and is similar to the 4PEG Model section. Motivational Design replaces the 4PEG Instructional Design, as shown in Figure 4, so that the five aspects that respond to most of the educational ratings are fully covered. Each aspect presents an assessment of the game components that stimulate the attention that players lend to the instructions [20–23].
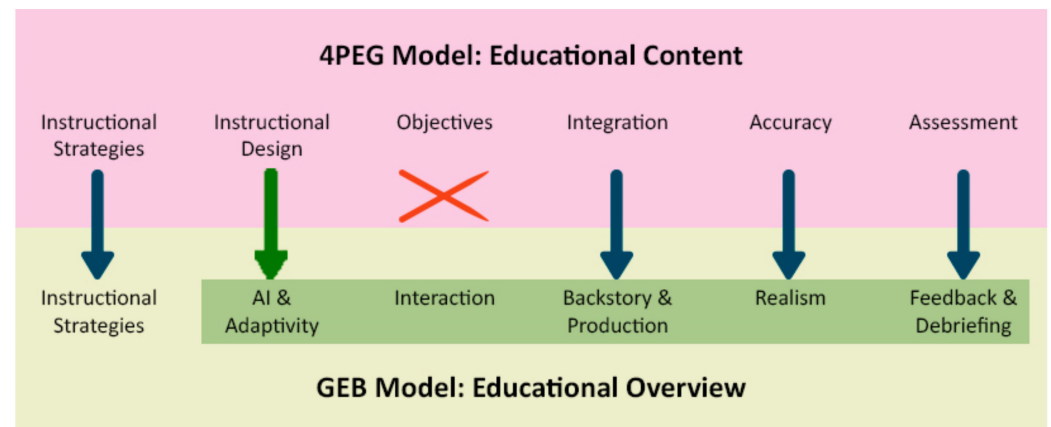

Figure 4. Adaption of 4PEG Model Educational Content to GEB model Educational Overview.

Instructional Strategies

Serious games are principally designed on the basis of constructivist theory [51,52]. This kind of instruction centers attention on the knowledge a person already possesses [10]. The game context therefore introduces the player to an unfamiliar environment that can be explored and where personal skills can be tested. The learning takes place while applying previous knowledge or acquiring new knowledge to resolve a conflict [53]. According to several authors [52,54–56], a very effective way to motivate this game confrontation is through the introduction of competition. Playing against other people engages the user to the gameplay. Nevertheless, if contests are not permissible due to the technology or the learning goals, a time trial will always be an optimal solution [3].

The Kaufman and Flanagan [18] embedded design for prosocial games is another strategy upon which the GEB model is based. This model could be applied to every serious game whose aim is to change behaviors or prejudices towards social matters. The authors

explain how to introduce social learning content to the players. Firstly, it has been shown in several works [57–59] that instruction on real-life social problems should not be directly approached. It is likely to trigger psychological defenses within the user that can block the persuasive aim or even inspire beliefs that are contrary to the intended outcomes. Embedded design is instead based on the premise that hiding educational social content within the game makes the player more receptive to the learning process [60].

There are many ways to insert social contents within embedded design. The first is to intermix non-focal educative and fun messages through mechanics and narratives. Additionally, it is demonstrated in this study that these kinds of content should be unbalanced. When playful topics are more numerous, instruction is better hidden so the learning results improve [24]. The second strategy is to disguise the true objective through a particular video game genre. Quizzes and puzzles are associated with educational content, less so adventure and platform games. Designing experiences in which the gameplay and the narrative stand out creates a safe perception for the player to accept the learning [18]. Finally, another way to hide the educational content is by distancing the narrative from real life issues [61]. Representing problems in a more abstract story, including fantasy elements, better immerses the player in the game, circumventing any possible negative reactions towards the learning process [47].

Motivational Design

This subsection evaluates how players are motivated to learn based on the Attention Relevance Confidence and Satisfaction (ARCS) Model [62]. This theory explains the main features that need stimulation so that users will focus their attention on the gameplay and achieve the educational aims. Attention activates player responses when educational incentives are offered [36]. Relevance is achieved when the game manages to relate the life experiences of a user to the learning content [37]. Confidence is built up, so that the player feels a positive expectation of their capability to complete the game tasks. If the player can prove both the knowledge and the skills that have been acquired, then satisfaction will also be felt in the final learning process [35].

Following the review of Ravyse et al. [3], Motivational Design is divided into Backstory and Production, Realism, Feedback and Debriefing, Artificial Intelligence (AI) and Adaptivity and Interaction. Their review of serious games literature summarized five main factors to improve educational impact. These factors are included in this subsection as they are equitable with the stimulus proposed in the ARCS model; some of them are connected to the 4PEG Model Educational Content characteristics. Backstory and Production are similar to the 4PEG Integration, in so far as these activities define how the learning content is embedded in the game. Realism is close to Accuracy because this feature relates to educational adaptation within the gameplay. Lastly, Feedback and Debriefing is comparable to Assessment, due to the focus on the player assuring the learning content. On the other hand, AI and Adaptivity and Interaction elements are totally new in this GEB model.

- Backstory and Production

Story is one of the main elements in any game, an essential part of the GEB model. This feature can be defined as the sequence of either linear, or branching or emergent events during the game [38]. Story is based on a main plot within which the game experience progresses and likewise the story gathers together all the strands in the narrative such as the environment and life events [63,64]. Each component engages the player, stimulating their discovery throughout the games that involve learning [65,66]. However, this knowledge must be appealing and integrated into the story without interrupting the fun [67]. Otherwise, players may become frustrated and if they stop playing, then the educational aim will not be achieved [68,69].

The conflict can be resolved by ensuring that the learning competences also progress throughout the game [65,66,70,71]. Moreover, linking the reward system (mechanics) to the desired educational outcomes will reinforce player engagement [72]. Another important aspect that motivates players is the ability to explore the story and as it develops choose

between options [67,73,74]. Knowing the player's desire to control the game, the developer team should adapt its possibilities to the learning style [75]. Some optimal tools would be to offer dialogues with multi-responses or scenario variations [3]. Finally, in addition to its attractiveness, the storyline must offer a fitting context to the educational material [51,65,76]. Otherwise, an extra-cognitive load will be added, decreasing the level of immersion. If reality cannot be reflected in the game, abstractive narrative may be necessary [25]. However, this variation will not be a problem, if it also serves to stimulate the player's intrinsic motivation to continue the game [3].

- Realism

Realism can be defined as fidelity to the physical, functional and psychological dimensions. In the game, it is represented by graphics, sounds and dynamics which correspond to the real world [47]. There is no consensus within academic circles over the importance of quality representation for learning. Nevertheless, a strong dependence on age is evident: children like artistic graphics and adults prefer physical fidelity [3]. Regarding this issue, various authors [47,76,77] support focusing the realist graphics, sounds and dynamics on the learning contents to prevent any distraction.

Realist aspects also involve stimulating a desire in the player to create a personal avatar [67,68] as a reflection of themselves [78]. It reinforces the user's sense of relevance in the game and immersive learning [79]. Although customization may be difficult to integrate in the development of a game, one advisable option is to offer a range of eligible characters with different appearances [3]. Another concern is the engagement with Non-Player Characters (NPC). The way the NPCs communicate and play different roles will affect the user's attention. Rather than textual dialogues, players prefer voice dialogues that introduce a heightened sense of realism [80]. People are accustomed to instructions communicated through intonation and facial expressions [81]. Therefore, user engagement is encouraged by showing NPC emotions and communicating through natural language and voice instead of text [23].

- AI and Adaptivity

Artificial Intelligence (AI) will adapt the game to the player's actions [3] through its agents both in an immediate and in a goal-directed way [82]. Adaptivity responds to the user's profile [83,84] and activity that is tracked and recorded on a database, activating programmatic flags to adapt to the choices [85]. This aspect is essential to guide players through the game and make the learning more effective [86–88]. The least intrusive adaptation is to use NPCs to provide the necessary feedback [76], so that the immersive experience will not decrease. Moreover, the sense of enjoyment is greater when these reactions turn out to be natural and customized [23,89]. Another aspect is the difficulty of adaptation that depends on the player's ability to avoid any frustration [90]. Once again, the development possibilities must be taken into account [91], nonetheless some multi-choice gameplay variations should be offered [92].

- Interaction

Playing games is based on Interaction. Players perform an action and the game responds accordingly through the audiovisual interface, instigating the next action. Complicated interfaces that frustrate players [72,93] and induce additional cognitive load [55,94] are not appropriate for serious games [79]. In various studies, it has been shown that the mechanics controlling the interface should be as direct and easy to learn as possible [52,85,95,96]. In this way, the interface should easily show both the status and the tools that are needed to progress and learn from the game [97]. Moreover, different experiences show that the game should include a tutorial level so that players can learn to handle the interface [67,81,98,99]. Subsequently, the game should progressively complicate the interactions [65,71] to challenge player skill levels [100]. In addition, collaborative playing through player interaction leads to successful learning. Sharing tactics and solutions to solve conflicts can also increase learning. This exchange can be implemented through chats

and voice communication, unless the application is impossible, in which case the sharing may take place during the debriefing [3].

- Feedback and Debriefing

Feedback is presented in two different ways. First, within the game through the NPC interactions and reward mechanics which produce an immediate cause–effect response [25,70,80]. The player experiences progress through these exchanges and becomes more competitive, wishing to unlock more content [72,83,101,102]. In the inclusion of feedback, it is important to consider the target audience to which the serious game is directed. If among these there are people with disabilities or special needs, designers must adapt games and learning content for them [103].

Second, the context in which the game is played is relevant, since it accepts the possibilities of feedback [104]. Real-time teacher support should be introduced, to enhance in-game feedback [105]. The ideal way would be to facilitate a teacher avatar that communicates with the players through an in-game chat [54,84]. Debriefing is the second way feedback can be offered; an informative after-game session with all the players [24]. Educational content may be discussed at the meeting, helping to process the knowledge and to consolidate it once acquired [106]. If the progress-tracking has been recorded, these data will be very useful to identify common trends and problems [3].

2.2.3. Overall Balance

Overall Balance is the last section of the GEB model. This metric reflects an evaluation of the interrelation between the Game Overview and the Educational Overview once both have been considered. The balance between the fun and the learning context gives important information to consider the serious game impact. This section particularly guarantees the proper distribution of all contents. Overall Balance is based on the 4PEG Magic Bullet Rating subsection [10]. Nevertheless, the GEB model rating is focused on only two features, as shown in Figure 5. The first one adapts the 4PEG Overall Balance content, but renames it according to the Learning and Fun Balance. This aspect evaluates equal distribution and impact of the gameplay and educational elements. The second feature preserves the Magic Bullet name: Can Learn vs. Must Learn. The metric is related to the types of learning and their introduction in the different game actions [25].
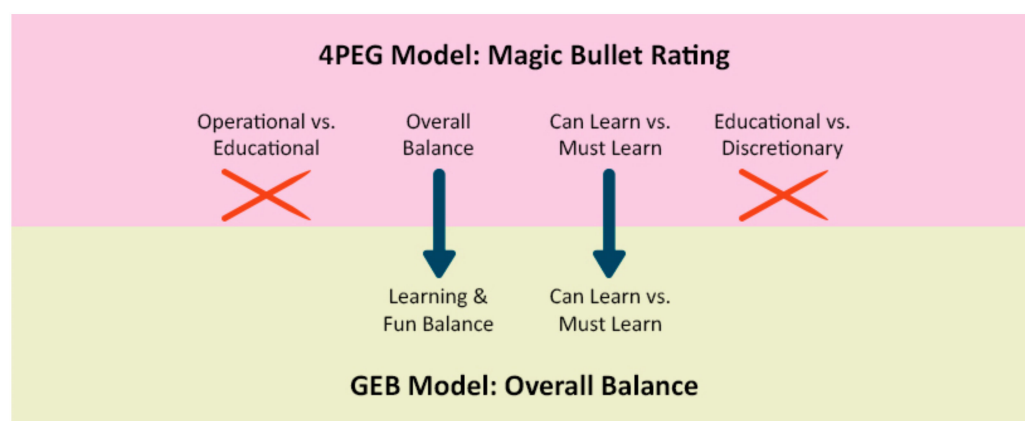

Figure 5. Adaption of 4PEG Model Magic Bullet Rating to GEB model Overall Balance.

Learning and Fun Balance

The balance between learning achievements and gaming is measured in this subsection. A perfectly balanced game will manage to apply most of the elements that were described in the Game Overview and the Educational Overview. As a result, the player will acquire the necessary knowledge and at the same time, enjoy all the gameplay interactions [10]. However, if the game cannot be considered fun in every interaction, the player must at least be engaged. For this purpose, narrative and mechanics will be applied, so that the

game evokes the curiosity, fantasy and challenge emotions that lead to control over the game [16]. The time taken to adapt to the learning and to the gameplay is another element that must also be considered [21,24]. The educational contents must therefore be equally distributed throughout the game [26]. Likewise, the controls that must be mastered will progressively increase in difficulty. Losing a good balance could lead the player to quit the game and never achieve the learning goals [27].

Can Learn vs. Must Learn

The Can Learn vs. Must Learn feature measures the distribution of the different kinds of knowledge throughout the game. On the one hand, the learning that the players must acquire should be embedded in the interactions that are necessary to complete the game. If the basic narrative and gameplay actions include these contents, then the game is educationally well designed [21]. In contrast, there is a type of learning that can be acquired, although it is not essential for the educational aim. Content inclusion of this sort responds to the player's curiosity to expand the context. The Can Learn knowledge should be included within the exploration possibilities of the game [25]. By no means should most of this content that might distract the player appear in the mandatory actions. In addition, a good serious game design should balance both types of learnings. The Can Learn content should not be more than the Must Learn content, otherwise the player will spend too much time on the game. Likewise, restricting the Can Learn context might leave the user unsatisfied [107].

2.2.4. Omitted Issues

The 4PEG model sections that were removed from the GEB model, rather than adapted to it, are discussed under this heading. The reasons for these decisions in favor of removal were based on the inclusion of these contents in other GEB features with pre-defined design needs and metrics. Keeping these aspects might have proved repetitive and confusing.

The characteristics erased from the Gameplay and Aesthetics section were:

- Content and Originality. Evaluated under this point are the appropriateness of the genre for the educational goal and the originality of the game compared to others. These characteristics have previously been explained in the GEB Instructional Strategies section.

- Game Progression. This paragraph measures the difficulty of transition throughout the gameplay. The aspect is included in AI and Adaptivity, belonging to the Motivational Design section.

- Settings and Characters. This feature relates to the adaptation of the game elements to the learning content and type of player. The Educational Overview evaluates these characteristics in the Backstory and Production and Realism guides.

The only non-adapted aspect in the 4PEG Educational Content section is the following:

- Objectives. This point does not correspond to any of the ARCS model features which are described in the Motivational Design subsection. On the contrary, this aspect is related to the explanation of the contents in the Instructional Strategies section. The objectives are used to evaluate the adequacy of the game for its purpose and its adaptability to various contexts.

The 4PEG Teacher Support section was completely removed from the GEB model. Becker acknowledged in his study that teacher intervention was not essential for a serious game. It would make little sense therefore to allocate 20% of the total rating to it in the GEB model. Likewise, the Feedback and Debriefing feature already deals with possible teacher support and the discussion of the acquired knowledge between players.

The following subsections that correspond to the 4PEG Magic Bullet Rating were also removed:

- Operational vs. Educational Learning: For the evaluation of learning, its evolution and its adaptation to the learning goals. These characteristics are evaluated under the GEB Instructional strategies and AI and Adaptivity sections.
- Educational vs. Discretionary Learning: An aspect that regards the ability to keep the player engaged with the experience to achieve the educational aim. These features are determined in the Backstory and Production process.

3. Results

3.1. GEB Metric Evaluation and Analysis Process

In this study, having defined the metrics, the three indie serious games designed with a social purpose in mind were all tested. These study cases are recent and well-known video games and have been acknowledged for raising awareness of various mental illnesses among players. Each narrative of the experience, its playability, and the mental illness it covers are explained in Supplementary File S2. The three games have also been chosen be-cause each case develops a different video game genre, so their adaptation to learning can be assessed. Furthermore, these are short experiences that volunteers could evaluate in a brief period of time. The survey results helped to explain which elements were to be measured in each section to choose the score. Moreover, these included personal questions on their hobbies for video games, their experience with each game, as well as any weak and strong points that they might like to highlight.

A total of 12 surveys of each game were performed with student volunteers from Burgos University, within an age range between 21 and 34 years old and with an equitable gender distribution. With this group, it was expected to have a spectrum of people with a greater or lesser habit of playing video games. This was because previous studies have determined that skilled players perceive learning better [108], while people less used to playing require more feedback and perceive the gameplay as more complicated [27].

The students played each game for at least an hour to appreciate the gameplay and educational content and to score its characteristics with the GEB metric. Results were entered in a supplementary Excel spreadsheet and Supplementary File S2 presents the analysis of the three games. First, one of the authors considered their gameplay, educational content and general balance to measure their suitability as educational games. In addition to the merits and weak points that were found, changes and improvements were suggested for serious games development in the future. Initially, the volunteers scores were interrelated with their personal variables of age and gender. No relationship was observed, so the analysis was carried out with the entire group of evaluators in each game. Following this, their results were reviewed and compared with the scores of the author, highlighting the similarities and differences, with accompanying explanations.

This metric falls within user research, so Sauro and Lewis [109] recommendations on the statistical study of the results were followed. The decisions made were based on the type of metric, which is a standardized questionnaire with a specific format and pre-established scores on a scale of 5. Its content validity originated from Section 2 through the review of the literature. On the other hand, its construct validity was ratified with statistical procedures to discover or confirm clusters of related items.

This is an experimental study that seeks to validate the usefulness of the metric and not the games on which it is applied. For this reason, there are certain statistical procedures that were discarded. Users have played the games and completed the entire metric, so the results for all features were valid and no confidence intervals or completion rates were needed. There was also no A/B test between different games because they do not require a competitive analysis and all users played at the same time, so it was not a variable to take into account. Lastly, they were not a continuous means on which a *t*-test could be performed between samples or users. The procedures that were performed and are explained in Section 3.2 were mean, deviation, correlation, regression and ANOVA.

Although some of the previously cited statistical analysis might not be useful in this study, some procedure should be followed in order to ensure that the concepts to

be evaluated in the metric which are closely related (e.g., Aesthetics with Realism or Technology and UE with AI and Adaptivity and Interaction), were not confused by the final users. To assure that the evaluated concepts were clearly explained in the GEBs metric questions and users were able to distinguish them (doing a proper evaluation) some statistical and machine-learning validation was performed on the results of the games evaluation and their results are collected in Section 3.3.

3.2. Analysis Results of GEB Metric

The analysis begins with the expert's evaluation of each of the games and their features, which can be found in Supplementary File S2. The scores between sections varied significantly, although the total result was close, 51 for Sym, 48 for Actual Sunlight and 60 for Neverending Nightmares. Regarding the different sections, Game Overview generated significant outcomes: different genres and development capacities might attract more players due to the gameplay. However, all three games failed the Educational Overview test, as Figure 6 shows. This result might have been expected, since these games are not developed by expert educators, although the stories introduce real experiences with a learning aim. In addition, none of these video games were designed within a workshop, so the Feedback and Debriefing score was low. Moreover, indie games also focus more on gameplay because the ultimate goal is selling. In any case, the Overall Balance, depending on the game, showed that if the player is motivated by the topic, then these tools are useful for learning. Likewise, applying the GEB metric proved useful for identifying the weaknesses which could be reinforced in similar serious games developing the same learning content or genre.
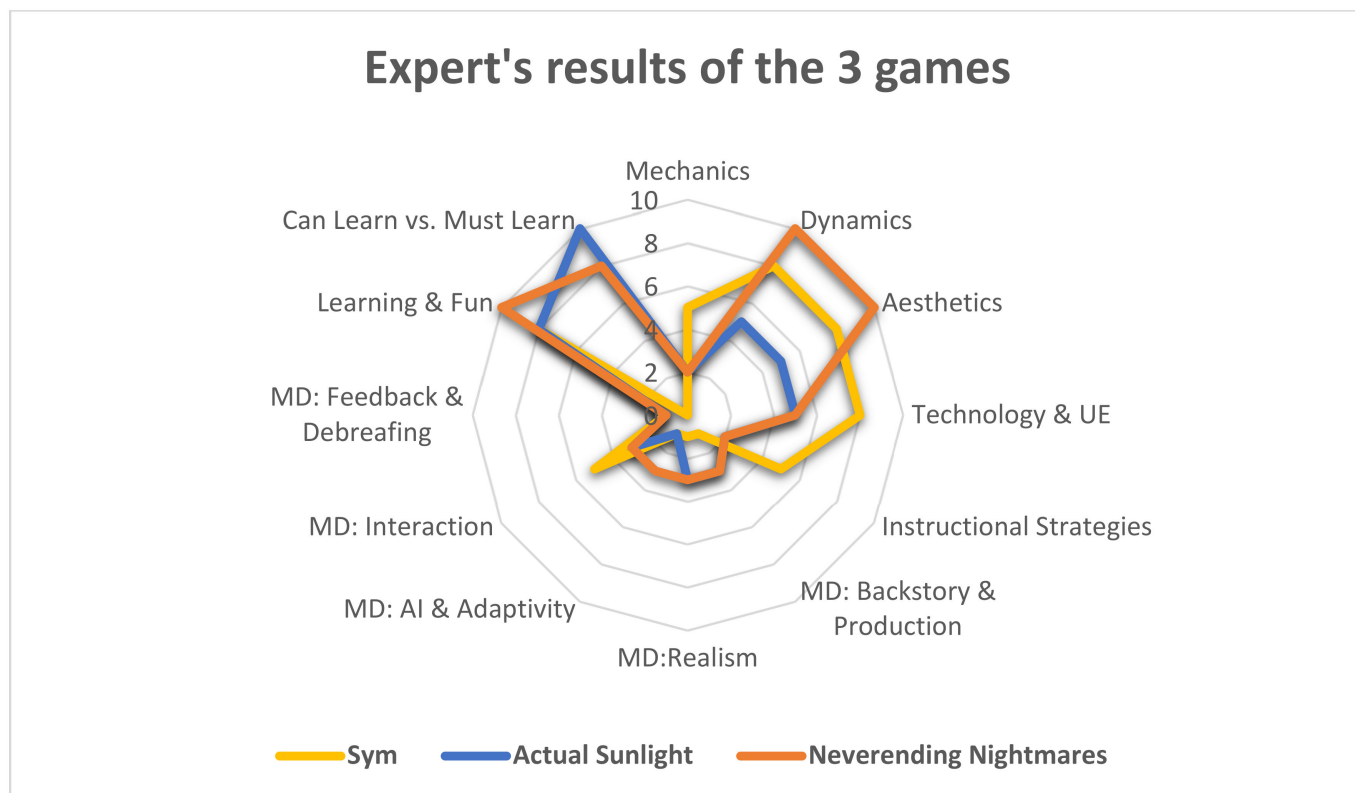

Figure 6. Expert's results of all 12 features from the three games.

Regarding the surveys and, since the sample for each game was small (12 users), the mean value is the most suitable to estimate the results average, because it has less error and bias than the median. The average scores attributed to both the respondents and the author were relatively similar, both in their sections and in most characteristics

as Supplementary File S2 explains in detail. This resemblance means that a user who experiences a game and rates it with this metric will be able to discern its educational merits and weaknesses. However, the standard deviations resulting from these surveys were quite high values, because the scores were within ranges. These data provide little information unless a feature is outside the set limits.

Nonetheless, it was evident from the results that the personal circumstances of each evaluator could affect the different scores. In the Game Overview section, mechanics score dropped when the gameplay was difficult (e.g., the platformer Sym game), as previous studies determined [27]. On the other hand, if the game genre did not challenge players, the dynamics score decreased. The Game Overview score can also fall when aesthetics did not respond to the evaluator's personal tastes regarding the graphic style, as the pixel art Actual Sunlight game showed. In the Educational Content section, the surveys showed that several evaluators never completely finished the games (since these required more than one hour). Therefore, Educational results determined the need for full access to the game's story in order to evaluate the inclusion of learning in the narrative. Likewise, respondents valued that the game genre was not associated with learning, such as the survival horror Neverending Nightmares. It was also seen that those players who appreciated the gameplay for their experience also valued the Educational Overview more, as other studies have highlighted [108]. In the Overall Balance section, the ease and habit of playing also affected the Learning and Fun feature. If the game is difficult or, on the contrary, too simple and repetitive, the mean drops. Finally, Can Learn vs. Must Learn feature depended heavily on the evaluators' playing time invested, interest and knowledge of each topic to distinguish between possible and mandatory educational contents.

3.3. Statistical and Machine-Learning Validation of the Games Evaluation

Finally, a statistical and machine-learning validation was performed on the results of the games evaluation to assure that those concepts are clearly explained in the GEBs metric questions and users are able to distinguish them (doing a proper evaluation). This evaluation included a violins plot analysis, a Pearson's correlation analysis, a Principal Components Analysis and an evaluation of the information and hidden patterns contained in the survey results by means of machine learning classifiers.

Firstly, a violins plot analysis was done for each game. Figure 7 shows this analysis. The violin plot shows the distribution of the users' answers for each variable. If two concepts were misunderstood or badly expressed in the GEBs metric questions, the answers might show almost the same distribution for both concept/variables. Therefore, if two concepts were misunderstood, their graphs should be very similar, game-by-game (e.g., a long one for Aesthetics and for Realism for Neverending Nightmares, a violin shape for Aesthetics and for Realism for Actual Sunlight and a round one for Aesthetics and for Realism for Sym); although they can be different if one game is compared with another (e.g., a long one for Neverending Nightmares and a violin shape for Sym).

Figure 7 shows very different shapes in most of the cases. Only Technology and UE with MD, Realism and Instructional Strategies with MD, Backstory and Production show the same similarities for the three games together. In the first case, both features evaluate audiovisual representation, but Technology and UE focuses on ease of control and mastery, and MD Realism on fidelity of representation and customization. In the second case, Instructional Strategies and Backstory and Production evaluate the integration of educational content into the game, but the former assesses the gameplay as a whole and the latter the plot. Like these, similarities are far from being significant, the violin analysis points to a clear difference between the proposed metric variables.
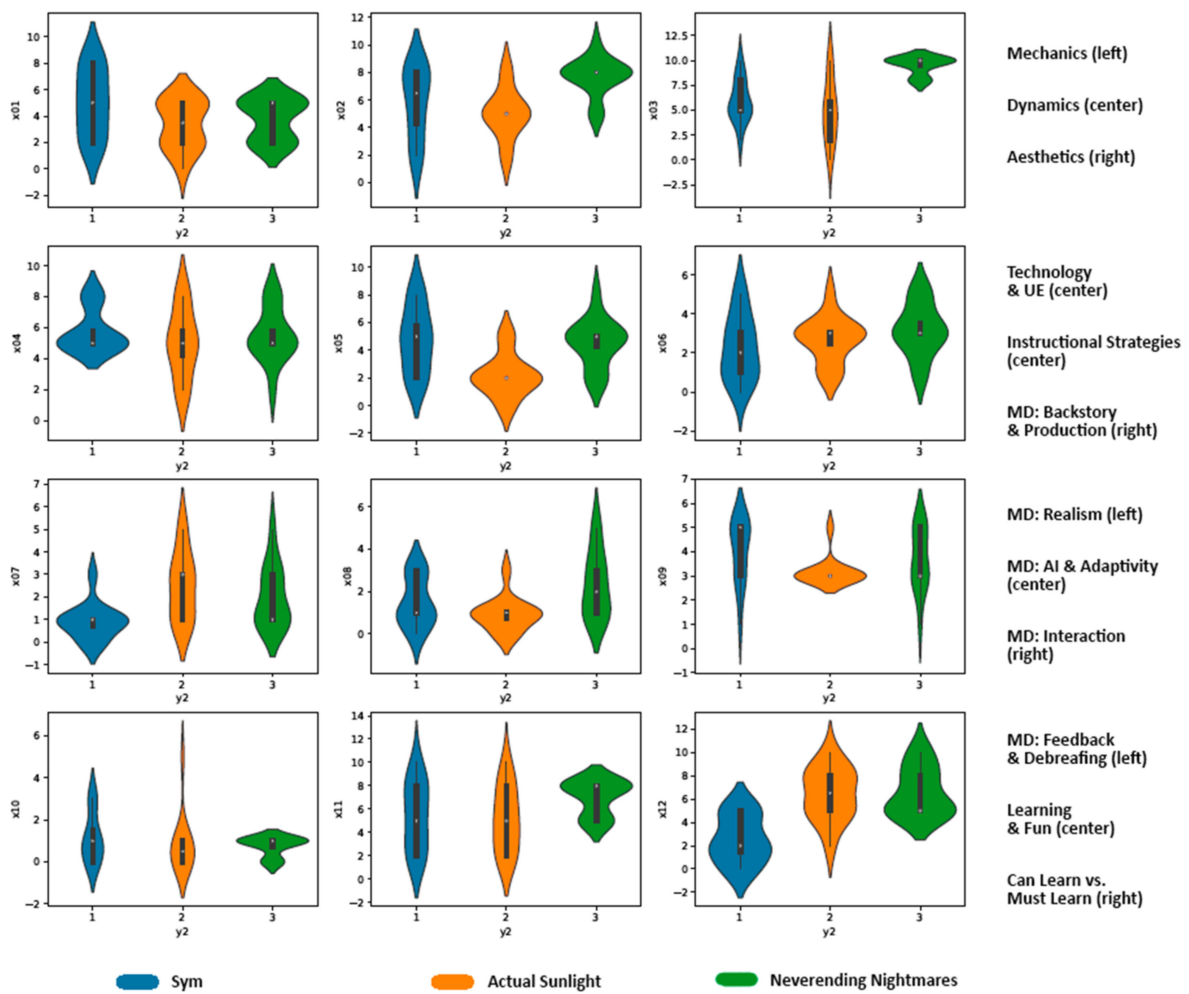
Figure 7. Violin plot analysis between the different model variables for each game.

Secondly, to reinforce the result of the violins analysis, a Pearson's correlation analysis between the different model variables for each game was performed. This analysis can determine if there is a statistical association between a couple of variables. Table 2 summarizes this analysis by collecting the *p*-values. The first value in each cell refers to Neverending Nightmares, the second to Actual Sunlight and the third to Sym. Statistical correlation between the two variables is found if the *p*-value is lower than 0.05. Therefore, *p*-values lower than this threshold are outlined in bold. The intention of this table is to show, from a first glance, that there are no cases where the three values are simultaneously in bold; a fact that would conclude that there is correlation between those two variables for the three games. An expanded version of this table can be found in the Excel supplementary File. There are some six correlations between two games: Backstory and Production (MD: B and P) and Feedback and Debriefing (MD: F and D) with AI and Adaptivity (MD: AI and A) and with Interaction (MD: I), and Instructional Strategies (IS) with Interaction (MD: I) and Aesthetics (A) with AI and Adaptivity (MD: AI and A). These results might be expected due to the nature of the datasets and the small sample size; users can evaluate two variables for one game in the same way (e.g., aesthetics and dynamics for Neverending Nightmares). However, this result does not imply a correlation between both variables in the metric; this correlation is only true if it appears for the three games at the same time (e.g., aesthetics and dynamics for the three games), because the diversity of the three evaluated games makes it impossible to find this correlation if its origin is not a badly defined metric (e.g., aesthetics and dynamics cannot be correlated for the three games, because they are not equally considered in the design of the three games).

Table 2. Pearson's correlation analysis between the different model variables for each game (first value in each cell refers to Sym, the second to Actual Sunlight and the third to Neverending Nightmares). *p*-values lower than 0.05 are outlined in bold.

	M	D	A	T and UE	IS	MD: B and P	MD: R	MD: AI and A	MD: I	MD: F and D	L and F	CL vs. ML
M	**0**	0.099 0.093 **0.045**	0.433 0.930 0.738	1.000 0.064 0.271	0.662 0.313 0.066	0.838 0.497 0.551	1.000 0.949 0.148	1.000 0.516 0.302	0.304 0.324 0.368	0.376 0.393 0.718	0.819 0.240 0.197	0.414 0.557 0.082
D	0.099 0.093 **0.045**	**0**	0.583 0.201 **0.001**	0.228 0.166 0.206	0.584 0.087 0.082	0.477 0.391 0.946	0.584 0.178 0.522	0.746 1.000 **0.033**	0.290 1.000 0.992	0.464 0.727 0.445	0.619 0.616 0.144	0.669 1.000 0.308
A	0.433 0.930 0.738	0.583 0.201 **0.001**	**0**	0.320 **0.037** 0.059	0.109 0.076 **0.005**	0.306 0.059 **0.020**	0.510 0.214 0.471	0.180 **0.040** **0.017**	0.199 0.081 0.101	0.200 0.222 0.936	**0.042** 0.326 **0.001**	0.523 0.134 0.161
T and UE	1.000 0.064 0.271	0.228 0.166 0.206	0.320 **0.037** 0.059	**0**	0.606 0.426 0.181	0.573 0.148 0.556	0.606 0.278 0.703	0.671 0.302 0.759	0.465 0.148 0.301	0.491 0.385 0.239	0.917 0.683 0.790	0.443 0.123 0.811
IS	0.662 0.313 0.066	0.584 0.087 0.082	0.109 0.076 **0.005**	0.606 0.426 0.181	**0**	0.051 0.054 0.150	0.319 0.694 0.094	0.130 **0.026** 0.959	0.151 **0.036** **0.015**	0.563 **0.028** 0.344	**0.007** 0.884 0.073	0.349 0.495 0.490
MD: B and P	0.838 0.497 0.551	0.477 0.391 0.946	0.306 0.059 **0.020**	0.573 0.148 0.556	0.051 0.054 0.150	**0**	0.275 0.717 0.555	**0.002** **0.045** 0.699	0.114 **0.022** **0.037**	0.210 0.066 0.912	0.194 0.068 **0.031**	0.896 0.479 0.168
MD: R	1.000 0.949 0.148	0.584 0.178 0.522	0.510 0.214 0.471	0.606 0.278 0.703	0.319 0.694 0.094	0.275 0.717 0.555	**0**	0.061 **0.042** 0.072	0.771 0.075 0.394	0.498 **0.047** 0.154	0.505 0.333 0.515	0.905 0.183 **0.005**
MD: AI and A	1.000 0.516 0.302	0.746 1.000 **0.033**	0.180 **0.040** **0.017**	0.671 0.302 0.759	0.130 **0.026** 0.959	**0.002** **0.045** 0.699	0.061 **0.042** 0.072	**0**	0.224 **0.001** 0.283	0.067 **0.001** **0.014**	0.100 0.667 0.878	0.544 0.474 0.540
MD: I	0.304 0.324 0.368	0.290 1.000 0.992	0.199 0.081 0.101	0.465 0.148 0.301	0.151 **0.036** **0.015**	0.114 **0.022** **0.037**	0.771 0.075 0.394	0.224 **0.001** 0.283	**0**	0.486 **0.001** **0.027**	0.151 0.356 0.152	0.568 0.587 0.794
MD: F and D	0.376 0.393 0.718	0.464 0.727 0.445	0.200 0.222 0.936	0.491 0.385 0.239	0.563 **0.028** 0.344	0.210 0.066 0.912	0.498 **0.047** 0.154	0.067 **0.001** **0.014**	0.486 **0.001** **0.027**	**0**	0.792 0.576 0.964	0.054 0.525 0.543
L and F	0.819 0.240 0.197	0.619 0.616 0.144	**0.042** 0.326 **0.001**	0.917 0.683 0.790	**0.007** 0.884 0.073	0.194 0.068 **0.031**	0.505 0.333 0.515	0.100 0.667 0.878	0.151 0.356 0.152	0.792 0.576 0.964	**0**	0.701 **0.023** 0.112
CL vs. ML	0.414 0.557 0.082	0.669 1.000 0.308	0.523 0.134 0.161	0.443 0.123 0.811	0.349 0.495 0.490	0.896 0.479 0.168	0.905 0.183 **0.005**	0.544 0.474 0.540	0.568 0.587 0.794	0.054 0.525 0.543	0.701 **0.023** 0.112	**0**

These two first analyses, shows that the features are well defined and explained in the metric, so that the volunteers have been able to evaluate clearly, and in an individual way, each metric element.

Thirdly, a Principal Components Analysis was done as a first attempt to verify if there was enough hidden information in the surveys to properly classify the three different videogames. Principal Components Analysis is a basic technique to order information, grouping in a linear way the original inputs to create new artificial inputs that maximize the information included in the dataset. Figure 8 shows the 2D representation of each user's survey for each game in terms of the two first principal components (the ones that have most of the extracted information from the dataset). Each colour refers to a different game. Figure 8 already points out that the GEB survey contains enough different information

to classify each game. Obviously, there are some users' surveys mixed up (lower right corner) when only those two first principal components are considered; however, as PCA is able to group the three games clearly, more complex classifiers will be able to extract more information from these datasets.
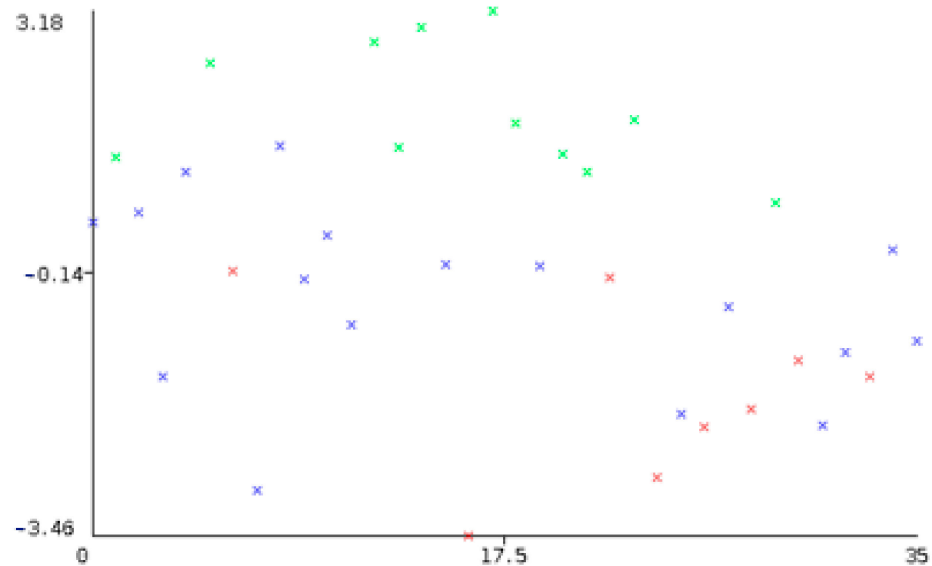

Figure 8. PCA representation for each game (different colours).

Fourthly, different machine learning classifiers were tested to evaluate if the survey results contain useful information. The Weka software tool was used for this task under a 10×10 validation scheme to assure the generalization of the obtained results. Machine learning algorithms are usually tested against a baseline method. The most common one is Näive-Bayes (called ZeroR in Weka), that always predicts the majority class. For this test, the game was considered as a class, or value to be predicted. Therefore, the class had three different possible values (1 for Neverending Nightmares, 2 for Actual Sunlight and 3 for Sym). The considered inputs were the 12 concepts evaluated in the GEB model plus a binary input about the level of expertise of the user in games (0: low experience in gaming, 1: high experience). The tested machine learning techniques were k-nearest neighbours (kNN), Decision Trees, Multilayer Perceptron and Random Forest. KNN is a very simple classifier that uses the closest example of any new data to build the prediction for the new one. Decision Trees are simple, unstable but visual classifiers. Multilayer Perceptions are able to predict complex relationships between inputs and outputs and are often used as a standard in machine learning tasks. Random Forest is an ensemble, a more recent machine learning technique, very insensitive to parameters tuning.

Table 3 summarizes the capability of those models to predict the evaluated game, considering the score given to a game by a user and his videogame playing habits. This capability was evaluated in terms of accuracy and the Receiver Operating Characteristic (ROC) Area. The accuracy is defined as the evaluations properly classified divided by the total of evaluations. The ROC area is the area under the Receiver Operating Characteristic curve and it is a more complex indicator, representing the ratio between the true positives and the false positives. Both quality indicators, accuracy and ROC area, are better the higher they are (accuracy close to 100% and ROC area close to 1.00). All the machine learning methods have been tested with their default values in Weka to simplify the tuning process, because a very high model accuracy was not the main objective of this analysis, but the level of hidden information in the dataset.

Table 3. Accuracy and ROC area for the machine-learning models.

Machine Learning Technique	Accuracy (%)	ROC Area
Näive-Bayes	27.78	0.389
KNN	61.1	0.722
Decision Trees	58.33	0.762
Multilayer Perceptron	75.00	0.983
Random Forest	75.00	0.924

Table 3 shows firstly that all machine learning techniques are statistically significantly better than the baseline method, the Näive-Bayes. Therefore, it can be assured that the GEBs surveys contain information different and coherent for each game, independent of the user who has filled in the survey. Secondly, this information is structured in a complex way, because more complex classifiers like artificial intelligence networks or ensembles achieve a better performance than more simple classifiers, like decision trees or kNNs. Thirdly, the hidden information cannot be extracted by just considering the answer from the previous user who gave the closest answer in the survey, because in this case kNN would be the most accurate method; this result is expected because the evaluation of a game by humans is a noisy task, and therefore many hidden and personal human factors might affect the evaluation of each user, making the use of the GEB metric a noisy task but also a predictable task. In this way, although users are different in their metrics values, there is coherence between their answers and the three games present a personal fingerprint that the machine learning algorithms can properly identify.

Considering these four different analyses, it can be concluded that the variables included in the GEB model and metric have been properly explained, the users can distinguish a clear difference when evaluating them in different games and the GEB survey can provide an individual fingerprint for each game that machine learning techniques can clearly inference.

4. Conclusions

A new metric, the Gaming Educational Balanced Model (GEB), has been proposed in this study to evaluate serious games. The theory is based on the MDA framework, adding the User Experience, to define the main gameplay sections. Likewise, the educational overview has been derived from the contents of 4PEG theory. However, each section of the GEB model has been adapted in the light of the latest research on serious games design [3,5,15,16,18,19,21–23]. These innovations on the necessary adaptability of the games to the learning pace of the players are incorporated in the metric. Finally, drawing from 4PEG theory, the importance of balancing the learning and the fun content has been customized in the GEB model, as well as the distribution of such content throughout the game experience.

As a novelty, a value system was created for this rubric in which the characteristics of gameplay and educational content were both worth 40%. The remaining 20% corresponded to the balance between fun and learning. Compared to the previous model, some points have been removed to make more general sections that cover every element in the game. The evaluation is made easier when the exact aspects are described and the same scores established for these sections [20,25,28,29]. The Game Overview subsections are Mechanics, Dynamics, Aesthetics and Technology and User Experience. The Educational Overview has been divided into Instructional strategies and Motivational design. This last subsection counts triple, because it has been divided into five educational features related to the learning adaptions. Lastly, Overall Balance splits the score between the distribution of Learning and Fun balance and Can Learn vs. Must Learn educational contents.

The rubric has also been applied to three indie games designed to raise awareness of various mental illnesses, in order to validate the GEB model. Each experience has been assessed by an author of this article and by 12 volunteers from the University of Burgos community. The results of both evaluations have shown that the GEB metric can be used

to determine the educational strengths and weaknesses of a serious game. However, the outcomes also showed that certain personal circumstances of the evaluator may affect the ratings. Therefore, the ideal user of this metric will be a player with knowledge and interest in the topics that are covered. Another possible option was to try each game with several people to mitigate these variable factors, as this paper has shown [27,108].

A statistical and machine-learning validation has been performed on the results of the games evaluation to assure that those concepts are clearly explained in the GEBs metric questions and users are able to distinguish them. This evaluation includes a violins plot analysis, a Pearson's correlation analysis, a Principal Components Analysis and an evaluation of the information and hidden patterns contained in the survey results by means of machine learning classifiers. The result of the analysis shows that the variables included in the GEB model and metric have been properly explained, and the users can distinguish a clear difference when evaluating them in different games and the GEB survey provides an individual fingerprint for each game that machine learning techniques can clearly inference.

It must be taken into account that the GEB model is a measure aimed at the most serious games that have assimilated the most important and general variables to assess its design. Therefore, in the evaluation of different games, factors such as the target audience or the topic must be considered. Learning contents cannot be introduced in the same way for adults as for children, just as they are attracted to different gameplays. Likewise, the game has to be adapted to the public with accessibility issues if it is addressed to them [5]. In the same way, the educational overview is valued differently when mental conditions are addressed as in the examples of this study [103], than if STEM education is taught [12,37]. As this article has explained, any evaluator could adapt the GEB metric features and percentages to their specific needs.

Future work will review what has been identified as more variable features to determine what changes in the rubric will decrease the subjectivity of the evaluator. From those variations, the GEB model will be applied to a wider variety of serious games, testing possible improvements for different audiences and topics. Likewise, although this paper has not observed any relationship between the evaluation results and the age and gender variables of the participants, its possible impact will also be studied. In addition, GEB will be used for the design of future serious games. The tests on the learning acquired with the developed videogames will also allow for the validation of the GEB theory.

References

1. Checa, D.; Bustillo, A. A review of immersive virtual reality serious games to enhance learning and training. *Multimed. Tools Appl.* **2020**, *79*, 5501–5527. [CrossRef]
2. Huang, W.; Roscoe, R.D.; Johnson-Glenberg, M.C.; Craig, S.D. Motivation, engagement, and performance across multiple virtual reality sessions and levels of immersion. *J. Comput. Assist. Learn.* **2021**, *37*, 745–758. [CrossRef]
3. Ravyse, W.S.; Seugnet Blignaut, A.; Leendertz, V.; Woolner, A. Success factors for serious games to enhance learning: A systematic review. *Virtual Real.* **2016**, *21*, 31–58. [CrossRef]
4. Lai, J.W.M.; Bower, M. Evaluation of technology use in education: Findings from a critical analysis of systematic literature reviews. *J. Comput. Assist. Learn.* **2019**, *36*, 241–259. [CrossRef]
5. Ferreira De Almeida, J.L.; dos Santos Machado, L. Design requirements for educational serious games with focus on player enjoyment. *Entertain. Comput.* **2021**, *38*, 100413. [CrossRef]
6. Mitgutsch, K.; Alvarado, N. Purposeful by design? A Serious Game Design Assessment Framework. In Proceedings of the International Conference on the Foundations of Digital Games—FDG'12, Raleigh, NC, USA, 29 May–1 June 2012.
7. Mayer, I. Towards a Comprehensive Methodology for the Research and Evaluation of Serious Games. *Procedia Comput. Sci.* **2012**, *15*, 233–247. [CrossRef]
8. Nasution, N.K.G.; Jin, X.; Singgih, I.K. Classifying games in container terminal logistics field: A systematic review. *Entertain. Comput.* **2022**, *40*, 100465. [CrossRef]
9. Hunicke, R.; Leblanc, M.; Zubek, R. MDA: A formal approach to game design and game research. In Proceedings of the AAAI Workshop on Challenges in Game AI, San José, CA, USA, 25–29 July 2004.
10. Becker, K. *Choosing and Using Digital Games in the Classroom: A Practical Guide*; Advances in Game-Based Learning; Springer International Publishing: Berlin/Heidelberg, Germany, 2017.
11. Martinez, K.; Menéndez-Menéndez, M.I.; Bustillo, A. Awareness, Prevention, Detection, and Therapy Applications for Depression and Anxiety in Serious Games for Children and Adolescents: Systematic Review. *JMIR Serious Games* **2021**, *9*, e30482. [CrossRef]
12. Dankov, Y.; Antonova, A.; Bontchev, B. Adopting User-Centered Design to Identify Assessment Metrics for Adaptive Video Games for Education. In *Interaction, Emerging Technologies and Future Systems V*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 289–297. [CrossRef]
13. Suryapranata, L.K.P.; Soewito, B.; Kusuma, G.P.; Gaol, F.L.; Warnars, H.L.H.S. Quality measurement for serious games. In Proceedings of the 2017 International Conference on Applied Computer and Communication Technologies (ComCom), Tumkur, Karnataka, India, 21–23 December 2017. [CrossRef]
14. Becker, K. Instructional ethology: Reverse engineering for serious design of educational games. In Proceedings of the Future Play, The International Conference on the Future of Game Design and Technology, Toronto, ON, Canada, 15–17 November 2007.
15. Junior, R.; Silva, F. Redefining the MDA Framework—The Pursuit of a Game Design Ontology. *Information* **2021**, *12*, 395. [CrossRef]
16. Kim, J.T.; Lee, W. Dynamical model for gamification of learning (DMGL). *Multimed. Tools Appl.* **2015**, *74*, 8483–8493. [CrossRef]
17. Winn, B.M. The Design, Play, and Experience Framework. In *Handbook of Research on Effective Electronic Gaming in Education*; IGI Global: Hershey, PA, USA, 2009; pp. 1010–1024. [CrossRef]
18. Kaufman, G.; Flanagan, M. A psychologically "embedded approach to designing games for prosocial causes. *Cyberpsychology J. Psychosoc. Res. Cyberspace* **2015**, *9*, 5. [CrossRef]
19. Pereira, M.; Winn, B.; Cerazotto, M.; Luiz, A.; Varella, P. Educational Digital Games: A theoretical framework about design models, learning theories and user experience. In Proceedings of the 7th International Conference, Design, User Experience and Usability, Theory and Practice, Las Vegas, NV, USA, 15–20 July 2018.
20. Caserman, P.; Hoffmann, K.; Müller, P.; Schaub, M.; Straßburg, K.; Wiemeyer, J.; Bruder, R.; Göbel, S. Quality Criteria for Serious Games: Serious Part, Game Part, and Balance. *JMIR Serious Games* **2020**, *8*, e19037. [CrossRef]
21. Toh, W.; Kirschner, D. Self-directed learning in video games, affordances and pedagogical implications for teaching and learning. *Comput. Educ.* **2020**, *154*, 103912. [CrossRef]
22. Lu, Y.L.; Lien, C.J. Are They Learning or Playing? Students' Perception Traits and Their Learning Self-Efficacy in a Game-Based Learning Environment. *J. Educ. Comput. Res.* **2019**, *57*, 1879–1909. [CrossRef]
23. Alexiou, A.; Schippers, M.C. Digital game elements, user experience and learning: A conceptual framework. *Educ. Inf. Technol.* **2018**, *23*, 2545–2567. [CrossRef]
24. Whitton, N.; Langan, M. Fun and games in higher education: An analysis of UK student perspectives. *Teach. High. Educ.* **2018**, *24*, 1000–1013. [CrossRef]
25. Greipl, S.; Moeller, K.; Ninaus, M. Potential and limits of game-based learning. *Int. J. Technol. Enhanc. Learn.* **2020**, *12*, 363. [CrossRef]
26. Namkoong, K.; Nah, S.; Record, R.A.; van Stee, S.K. Communication, Reasoning, and Planned Behaviors: Unveiling the Effect of Interactive Communication in an Anti-Smoking Social Media Campaign. *Health Commun.* **2016**, *32*, 41–50. [CrossRef]
27. Yang, J.C.; Chen, S.Y. An investigation of game behavior in the context of digital game-based learning: An individual difference perspective. *Comput. Hum. Behav.* **2020**, *112*, 106432. [CrossRef]
28. Connolly, T.M.; Boyle, E.A.; MacArthur, E.; Hainey, T.; Boyle, J.M. A systematic literature review of empirical evidence on computer games and serious games. *Comput. Educ.* **2012**, *59*, 661–686. [CrossRef]

29. Jacobs, R.S. Serious games: Play for change. In *The Video Game Debate 2*; Kowert, R., Quandt, T., Eds.; Routledge: London, UK, 2020; pp. 19–40. [CrossRef]

30. Rosenthal, S.; Ratan, R.A. Balancing learning and enjoyment in serious games: Kerbal Space Program and the communication mediation model. *Comput. Educ.* **2022**, *182*, 104480. [CrossRef]

31. Czauderna, A.; Guardiola, E. The Gameplay Loop Methodology as a Tool for Educational Game Design. *Electron. J. e-Learn.* **2019**, *17*, 207–221. [CrossRef]

32. Moizer, J.; Lean, J.; Dell'Aquila, E.; Walsh, P.; Keary, A.A.; O'Byrne, D.; di Ferdinando, A.; Miglino, O.; Friedrich, R.; Asperges, R.; et al. An approach to evaluating the user experience of serious games. *Comput. Educ.* **2019**, *136*, 141–151. [CrossRef]

33. Hamari, J.; Shernoff, D.J.; Rowe, E.; Coller, B.; Asbell-Clarke, J.; Edwards, T. Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Comput. Hum. Behav.* **2016**, *54*, 170–179. [CrossRef]

34. Sharma, T.G.; Hamari, J.; Kesharwani, A.; Tak, P. Understanding continuance intention to play online games: Roles of self-expressiveness, self-congruity, self-efficacy, and perceived risk. *Behav. Inf. Technol.* **2020**, *41*, 348–364. [CrossRef]

35. Chen, C.C.; Tu, H.Y. The Effect of Digital Game-Based Learning on Learning Motivation and Performance under Social Cognitive Theory and Entrepreneurial Thinking. *Front. Psychol.* **2021**, *12*, 750711. [CrossRef]

36. Sanchez, D.R.; Nelson, T.; Kraiger, K.; Weiner, E.; Lu, Y.; Schnall, J. Defining motivation in video game-based training: Exploring the differences between measures of motivation. *Int. J. Train. Dev.* **2021**, *26*, 1–28. [CrossRef]

37. Benton, L.; Vasalou, A.; Barendregt, W.; Bunting, L.; Révész, A. What's Missing: The Role of Instructional Design in Children's Games-Based Learning. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019. [CrossRef]

38. Alves, F. *Gamification: Como Criar Experiências de Aprendizagem Engajadoras: Um Guía Completo: Do conceito à Prática*, 2nd ed.; DVS Editora: São Paulo, Brazil, 2015.

39. Buley, L. *The User Experience Team of One*; Roselfeld: New York, NY, USA, 2013.

40. Garrett, J.J. *The Elements of User Experience: User-Centered Design for the Web and Beyond*, 2nd ed.; New Riders: Berkeley, CA, USA, 2011.

41. Malone, T.W. *What Makes Things Fun to Learn? A Study of Intrinsically Motivating Computer Games*; Palo Alto Research Center (Xerox): Palo Alto, CA, USA, 1980.

42. Malone, T.W.; Lepper, M.R. Making learning fun: A taxonomy of intrinsic motivations for learning. In *Aptitude, Learning and Instruction III: Conative and Affective Process Analyses*; Snow, R.E., Farr, M.J., Eds.; Erlbaum: Hillsdale, NJ, USA, 1987; pp. 223–253.

43. Biles, M.L.; Plass, J.L.; Homer, B.D. Designing Digital Badges for Educational Games. In *Learning and Performance Assessment: Concepts, Methodologies, Tools, and Applications*; IGI Global: Hershey, PA, USA, 2020; pp. 1349–1369. [CrossRef]

44. Ferrara, J. *Playful Design: Creating Game Experiences in Everyday Interfaces*; Rosenfeld Media: New York, NY, USA, 2012.

45. Miceli, M.; Castelfranchi, C. Meta-emotions and the complexity of human emotional experience. *New Ideas Psychol.* **2019**, *55*, 42–49. [CrossRef]

46. Bailen, N.H.; Wu, H.; Thompson, R.J. Meta-emotions in daily life: Associations with emotional awareness and depression. *Emotion* **2019**, *19*, 776–787. [CrossRef]

47. Fokides, E.; Atsikpasi, P.; Kaimara, P.; Deliyannis, I. Factors Influencing the Subjective Learning Effectiveness of Serious Games. *J. Inf. Technol. Educ. Res.* **2019**, *18*, 437–466. [CrossRef]

48. Menezes, F.M.; Silva, I.C.S.; Frosi, F.O. Game User Experience (UX): Explorando a Teoria da Diegese. In Proceedings of the SBGames 2017 (XVI SBGames (SBC)), São Paulo, SP, Brazil, 2–4 November 2017.

49. Buil, I.; Catalán, S.; Martínez, E. Exploring students' flow experiences in business simulation games. *J. Comput. Assist. Learn.* **2018**, *34*, 183–192. [CrossRef]

50. Scholtz, B.; Raga, L.; Baxter, G. Design and Evaluation of a «Gamified» System for Improving Career Knowledge in Computing Sciences. *Afr. J. Inf. Commun. (AJIC)* **2016**, *2016*, 7–32. [CrossRef]

51. Cheng, M.-T.; Su, T.; Huang, W.-Y.; Chen, J.-H. An educational game for learning human immunology: What do students learn and how do they perceive? *Br. J. Educ. Technol.* **2014**, *45*, 820–833. [CrossRef]

52. Zin, N.; Yue, W.S. Design and evaluation of history Digital Game Based Learning (DGBL) software. *J. Next Gener. Inf. Technol.* **2013**, *4*, 9–24. [CrossRef]

53. Jonassen, D.H. Thinking technology: Toward a constructivist design model. *Educ. Technol.* **1994**, *34*, 34–37.

54. Barab, S.; Pettyjohn, P.; Gresalfi, M.; Volk, C.; Solomou, M. Game-based curriculum and transformational play: Designing to meaningfully positioning person, content, and context. *Comput. Educ.* **2012**, *58*, 518–533. [CrossRef]

55. Hong, J.-C.; Tsai, C.-M.; Ho, Y.-J.; Hwang, M.-Y.; Wu, C.-J. A comparative study of the learning effectiveness of a blended and embodied interactive video game for kindergarten students. *Interact. Learn. Environ.* **2013**, *21*, 39–53. [CrossRef]

56. Hwang, G.-J.; Wu, P.-H.; Chen, C.-C. An online game approach for improving students' learning performance in web-based problem-solving activities. *Comput. Educ.* **2012**, *59*, 1246–1256. [CrossRef]

57. Dillard, J.P.; Shen, L. On the nature of reactance and its role in persuasive health communication. *Commun. Monogr.* **2005**, *72*, 144–168. [CrossRef]

58. Grandpre, J.; Alvaro, E.M.; Burgoon, M.; Miller, C.H.; Hall, J.R. Adolescent reactance and anti-smoking campaigns: A theoretical approach. *Health Commun.* **2003**, *15*, 349–366. [CrossRef]

59. Quick, B.L.; Considine, J.R. Examining the use of forceful language when designing exercise persuasive messages for adults: A test of conceptualizing reactance arousal as a two-step process. *Health Commun.* **2008**, *23*, 483–491. [CrossRef]

60. Kaufman, G.; Flanagan, M.; Seidman, M. Creating stealth game interventions for attitude and behavior change: An "Embedded Design" model. In Proceedings of the Digital Games Research Association (DiGRA) Conference, Lüneburg, Germany, 14–17 May 2015.

61. Kaufman, G.F.; Flanagan, M. Lost in translation: Comparing the impact of an analog and digital version of a public health game on players' perceptions, attitudes, and cognitions. *Int. J. Gaming Comput.-Mediat. Simul. (IJGCMS)* **2013**, *5*, 1–9. [CrossRef]

62. Keller, J.M. Strategies for stimulating the motivation to learn. *Perform. Instr.* **1987**, *26*, 1–7. [CrossRef]

63. Shute, V.; Rahimi, S.; Smith, G.; Ke, F.; Almond, R.; Dai, C.; Kuba, R.; Liu, Z.; Yang, X.; Sun, C. Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *J. Comput. Assist. Learn.* **2020**, *37*, 127–141. [CrossRef]

64. Sun, C.T.; Chen, L.X.; Chu, H.M. Associations among scaffold presentation, reward mechanisms and problem-solving behaviors in game play. *Comput. Educ.* **2018**, *119*, 95–111. [CrossRef]

65. Couceiro, R.M.; Papastergiou, M.; Kordaki, M.; Veloso, A.I. Design and evaluation of a computer game for the learning of Information and Communication Technologies (ICT) concepts by physical education and sport science students. *Educ. Inf. Technol.* **2013**, *18*, 531–554. [CrossRef]

66. Hämäläinen, R. Designing and evaluating collaboration in a virtual game environment for vocational learning. *Comput. Educ.* **2008**, *50*, 98–109. [CrossRef]

67. Ke, F.; Abras, T. Games for engaged learning of middle school children with special learning needs. *Br. J. Educ. Technol.* **2013**, *44*, 225–242. [CrossRef]

68. Brom, C.; Sisler, V.; Slavik, R. Implementing digital game-based learning in schools: Augmented learning environment of 'Europe 2045'. *Multimed. Syst.* **2010**, *16*, 23–41. [CrossRef]

69. Hwang, G.-J.; Yang, L.-H.; Wang, S.-Y. A concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Comput. Educ.* **2013**, *69*, 121–130. [CrossRef]

70. Cheng, M.-T.; Annetta, L. Students' learning outcomes and learning experiences through playing a serious educational game. *J. Biol. Educ.* **2012**, *46*, 203–213. [CrossRef]

71. Ke, F. A case study of computer gaming for math: Engaged learning from gameplay? *Comput. Educ.* **2008**, *51*, 1609–1620. [CrossRef]

72. Kiili, K. Content creation challenges and flow experience in educational games: The IT-Emperor case. *Internet High. Educ.* **2005**, *8*, 183–198. [CrossRef]

73. Kickmeier-Rust, M.D.; Albert, D. Micro-adaptivity: Protecting immersion in didactically adaptive digital educational games. *J. Comput. Assist. Learn.* **2010**, *26*, 95–105. [CrossRef]

74. Verpoorten, D.; Castaigne, J.-L.; Westera, W.; Specht, M. A quest for meta-learning gains in a physics serious game. *Educ. Inf. Technol.* **2014**, *19*, 361–374. [CrossRef]

75. Westera, W. Why and How Serious Games can Become Far More Effective: Accommodating Productive Learning Experiences, Learner Motivation and the Monitoring of Learning Gains. *J. Educ. Technol. Soc.* **2019**, *22*, 59–69.

76. Bellotti, F.; Berta, R.; De Gloria, A.; Primavera, L. Enhancing the educational value of video games. *Comput. Entertain.* **2009**, *7*, 1–18. [CrossRef]

77. Annetta, L.A.; Minogue, J.; Holmes, S.Y.; Cheng, M.-T. Investigating the impact of video games on high school students' engagement and learning about genetics. *Comput. Educ.* **2009**, *53*, 74–85. [CrossRef]

78. Gee, J.P. Good video games and good learning. *Phi Kappa Phi Forum* **2005**, *85*, 33–37.

79. González-González, C.; Blanco-Izquierdo, F. Designing social videogames for educational uses. *Comput. Educ.* **2012**, *58*, 250–262. [CrossRef]

80. Johnson, C.I.; Mayer, R.E. Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Comput. Hum. Behav.* **2010**, *26*, 1246–1252. [CrossRef]

81. Van Eck, R. The effect of contextual pedagogical advisement and competition on middle-school students' attitude toward mathematics and mathematics instruction using a computer-based simulation game. *J. Comput. Math. Sci. Teach.* **2006**, *25*, 165–195.

82. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Pearson New International Edition: Harlow, UK, 2014.

83. Cheng, M.-T.; Lin, Y.-W.; She, H.-C. Learning through playing Virtual Age: Exploring the interactions among student concept learning, gaming performance, in-game behaviors, and the use of in-game characters. *Comput. Educ.* **2015**, *86*, 18–29. [CrossRef]

84. González-González, C.; Toledo-Delgado, P.; Collazos-Ordóñez, C.; González-Sánchez, J. Design and analysis of collaborative interactions in social educational videogames. *Comput. Hum. Behav.* **2014**, *31*, 602–611. [CrossRef]

85. Soflano, M.; Connolly, T.; Hainey, T. An application of adaptive games-based learning based on learning style to teach SQL. *Comput. Educ.* **2015**, *86*, 192–211. [CrossRef]

86. Vanbecelaere, S.; Van den Berghe, K.; Cornillie, F.; Sasanguie, D.; Reynvoet, B.; Depaepe, F. The effectiveness of adaptive versus non-adaptive learning with digital educational games. *J. Comput. Assist. Learn.* **2019**, *36*, 502–513. [CrossRef]

87. Bontchev, B.; Georgieva, O. Playing style recognition through an adaptive video game. *Comput. Hum. Behav.* **2018**, *82*, 136–147. [CrossRef]

88. Jagušt, T.; Botički, I.; So, H.J. Examining competitive, collaborative and adaptive gamification in young learners' math learning. *Comput. Educ.* **2018**, *125*, 444–457. [CrossRef]

89. Thompson, D.; Baranowski, T.; Buday, R.; Baranowski, J.; Thompson, V.; Jago, R.; Griffith, M. Serious video games for health: How behavioral science guided the development of a serious video game. *Simul. Gaming* **2010**, *41*, 587–606. [CrossRef]

90. Bellotti, F.; Berta, R.; De Gloria, A.; D'Ursi, A.; Fiore, V. A serious game model for cultural heritage. *J. Comput. Cult. Herit.* **2012**, *5*, 1–27. [CrossRef]

91. Ketamo, H.; Kiili, K. Conceptual change takes time: Game based learning cannot be only supplementary amusement. *J. Educ. Multimed. Hypermedia* **2010**, *19*, 399–419.

92. Virvou, M.; Katsionis, G.; Manos, K. Combining software games with education: Evaluation of its educational effectiveness. *J. Educ. Technol. Soc.* **2005**, *8*, 54–65.

93. Van der Spek, E.D.; Van Oostendorp, H.; Meyer, J.J.C. Introducing surprising events can stimulate deep learning in a serious game. *Br. J. Educ. Technol.* **2013**, *44*, 156–169. [CrossRef]

94. Hwang, G.-J.; Chiu, L.-Y.; Chen, C.-H. A contextual game-based learning approach to improving students' inquiry-based learning performance in social studies courses. *Comput. Educ.* **2015**, *81*, 13–25. [CrossRef]

95. Chittaro, L.; Buttussi, F. Assessing knowledge retention of an immersive serious game vs. a traditional education method in aviation safety. *IEEE Trans. Vis. Comput. Graph.* **2015**, *21*, 529–538. [CrossRef] [PubMed]

96. Knight, J.; Carley, S.; Tregunna, B.; Jarvis, S.; Smithies, R.; de Freitas, S.; Dunwell, I.; Mackway-Jones, K. Serious gaming technology in major incident triage training: A pragmatic controlled trial. *Resuscitation* **2010**, *81*, 1175–1179. [CrossRef]

97. Verma, V.; Baron, T.; Bansal, A.; Amresh, A. Emerging practices in game-based assessment. In *Game-Based Assessment Revisited*; Ifenthaler, D., Kim, Y.J., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 327–346. [CrossRef]

98. Hämäläinen, R. Using a game environment to foster collaborative learning: A design-based study Technology. *Pedagog. Educ.* **2011**, *20*, 61–78. [CrossRef]

99. Hwang, G.-J.; Sung, H.-Y.; Hung, C.-M.; Yang, L.-H.; Huang, I. A knowledge engineering approach to developing educational computer games for improving students' differentiating knowledge. *Br. J. Educ. Technol.* **2013**, *44*, 183–196. [CrossRef]

100. Csikszentmihalyi, M. *Flow: The Psychology of Optimal Experience*, 2nd ed.; Harper Perennial: New York, NY, USA, 2008.

101. Kuk, K.; Milentijević, I.; Rančić, D.; Spalević, P. Pedagogical agent in Multimedia Interactive Modules for Learning—MIMLE. *Expert Syst. Appl.* **2012**, *39*, 8051–8058. [CrossRef]

102. Wilson, A.S.; Broadbent, C.; McGrath, B.; Prescott, J. Factors Associated with Player Satisfaction and Educational Value of Serious Games. In *Serious Games and Edutainment Applications*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 513–535. [CrossRef]

103. Francillette, Y.; Boucher, E.; Bouchard, B.; Bouchard, K.; Gaboury, S. Serious games for people with mental disorders: State of the art of practices to maintain engagement and accessibility. *Entertain. Comput.* **2021**, *37*, 100396. [CrossRef]

104. Lugmayr, A.; Sutinen, E.; Suhonen, J.; Sedano, C.I.; Hlavacs, H.; Montero, C.S. Serious storytelling—A first definition and review. *Multimed. Tools Appl.* **2016**, *76*, 15707–15733. [CrossRef]

105. ter Vrugte, J.; de Jong, T.; Vandercruysse, S.; Wouters, P.; van Oostendorp, H.; Elen, J. Computer game-based mathematics education: Embedded faded worked examples facilitate knowledge acquisition. *Learn. Instr.* **2017**, *50*, 44–53. [CrossRef]

106. Crookall, D. Engaging (in) gameplay and (in) debriefing. *Simul. Gaming* **2014**, *45*, 416–427. [CrossRef]

107. Rahimi, S.; Shute, V.; Kuba, R.; Dai, C.P.; Yang, X.; Smith, G.; Alonso Fernández, C. The use and effects of incentive systems on learning and performance in educational games. *Comput. Educ.* **2021**, *165*, 104135. [CrossRef]

108. Darvishi, M.; Seif, M.; Sarmadi, M.; Farajollahi, M. An investigation into the Factors Affecting Perceived Enjoyment of Learning in Augmented Reality: A Path Analysis. *Interdiscip. J. Virtual Learn. Med. Sci.* **2020**, *11*, 224–235. [CrossRef]

109. Sauro, J.; Lewis, J.R. *Quantifying the User Experience*; Elsevier Gezondheidszorg: Amsterdam, The Netherlands, 2016.