A methodology to assess the effectiveness of serious games and infer player learning outcomes

Ángel Serrano-Laguna[1], Borja Manero[1], Manuel Freire[1], Baltasar Fernández-Manjón[1]

[1] Department of Software Engineering and Artificial Intelligence, Complutense University of Madrid. Madrid (Spain)

Corresponding Author:
Ángel Serrano-Laguna.

Department of Software Engineering and Artificial Intelligence
Complutense University of Madrid
Facultad de Informática
C. Profesor José García Santesmases, s/n.
28040 Madrid (Spain)

Email: angel.serrano@fdi.ucm.es

Abstract

Although serious games are proven educational tools in many educational domains, they lack reliable, automated and repeatable methodologies to measure their effectiveness: what do players know after playing a serious game? Did they learn with it? Literature research shows that the vast majority of serious games are assessed through questionnaires, which strikes a stark contrast with current trends in the video game industry. Commercial videogames have been learning from their players through Game Analytics for years, using non-disruptive game tracking. In this paper, we propose a methodology to assess serious games effectiveness using non-disruptive in-game tracking. The methodology proposes a design pattern that structures the delivery of educational goals within the game. This structure also allows inferring learning outcomes for each individual player, which, when aggregated, would determine the effectiveness of the serious game. We have tested this methodology with a serious game that was played by 320 students. The proposed methodology allowed us to infer players' learning outcomes and assess game effectiveness and to spot issues in the game design.

This research has been partially financed by the Ministry of Education, Culture and Sport of Spain, through their FPU Programme (grant FPU12/04310), by the Regional Government of Madrid (eMadrid S2013/ICE-2715), by the Complutense University of Madrid (GR3/14-921340), by the Ministry of Education (TIN2013-46149-C2-1-R), by the RIURE Network (CYTED 513RT0471) and by the European Commission (RAGE H2020-ICT-2014-1-644187, BEACONING H2020-ICT-2015-687676).

Keywords

serious games, learning analytics, game design, learning outcomes analysis, educational games

1. Introduction

A serious game is a video game designed with purposes beyond pure entertainment [1]. Serious games are multimedia tools by nature. As a subfamily of videogames, they combine different types of media (animations, music, text…) to create immersive experiences for the players. Their versatility allow them to be used as tools with many applications in different domains. One of the main ones is education, where they have become proven learning tools: they are used across many domains with multiple goals and formats, and their acceptance and effectiveness is almost always positive [2, 3]. Traditionally, a large percentage of serious games has been both developed and deployed by educational researchers, limiting their scope and reach. This trend is starting to change. Nowadays, widespread use of Virtual Learning Environments (VLE) allows for the application of serious games in unprecedented scales. To reach their full potential, serious games should adopt the latest advances in education and commercial videogames [4].

On-line education has increased exponentially in recent years, and many students now learn through Internet-connected devices. This vastly increases the amount of educational data available for analysis. Disciplines such as Learning Analytics (LA) or Educational Data Mining (EDM) study the patterns inside students' interactions to better understand the underlying learning processes [5, 6]. This knowledge can be used by different stakeholders with diverse purposes: from university administrators calculating dropout rates in each class, to teachers identifying students at risk of school failure [7].

Serious games (and video games in general) are particularly well suited for data analysis. Their highly interactive nature, based on a constant loop of user input followed by game feedback, pose them as rich sources of interaction data. These interactions can be later analyzed to explore how users play, and, in the case of serious games, understand how users learn.

The video game industry has been performing these types of analysis in commercial games for years, via Game Analytics (GA) [8]. One of the main uses of GA is to measure balance in gameplay: a balanced video game is one that keeps its players in the flow zone, a state where the player feels challenged by the game, but neither bored nor frustrated [9]. GA helps to locate parts inside games where players get stuck or quit; and moments where a game's mechanics or internal rules fall short. GA also provides clues on how to fix these problems.

Commercial video games usually collect data from their players in a non-disruptive way, with tracking systems that go unnoticed by the players [10]. However, according to the literature [11], the main method to assess any aspect of a serious game is the use of questionnaires filled by players. There is a clear need to combine the emerging disciplines of LA and EDM with the non-disruptive techniques of GA to provide reliable, automated and repeatable assessment for serious games.

Serious game assessment can focus on many results, such as usability, engagement or motivation. However, learning outcomes is the result most stakeholders want to obtain from serious games [12]. Learning outcomes are also the most frequent result assessed in recent serious games [11], and

some authors even believe that such outcomes could be used to replace standardized tests [13]. However, multiple issues with serious games must first be addressed. One of them is the lack of methods to assess serious games effectiveness [14]: teachers, lecturers and policy-makers need guarantees that serious games are effective enough to be used in the classroom. In this regard, the use of GA techniques with serious games can provide stakeholders with objective and reliable data.

In this paper, we propose a methodology to infer learning outcomes and serious games effectiveness based on non-disruptive tracking. The methodology targets two different phases in the life of a serious game: 1) its design and implementation, where we propose a game-design pattern to shape the delivery of the educational content throughout the game, and 2) its validation and deployment, where we propose an analysis, based on the game-design pattern, to infer learning outcomes and game effectiveness.

The paper is structured as follows: Section 2 presents a research review on serious game assessment. Section 3 presents the methodology and section 4 describes an experimental case study where the methodology was applied. Section 5 presents the results of the case study, which are then discussed in Section 6. Finally, section 7 presents the conclusions, some limitations, and future work.

2. Serious games assessment

Although the most common tool to assess serious games are questionnaires [11], several authors have addressed the implications of using non-disruptive tracking for this task. Authors have proposed a set of minimum requirements to enable automatic assessment in serious games [15], and have addressed the game design implications of combining learning analytics and serious games [16]. The project ADAGE [17] is a framework that defines several “assessment mechanics” that capture basic gameplay progression and critical achievements. Similarly, we have previously proposed a set of universal “traces”, particularly interesting for serious games, that can be emitted by any video game [18].

Other authors have implemented their own *ad-hoc* analytics to, for instance, analyze players’ steps in a math puzzle to predict their movements based on current game state [19], assess learning outcomes analyzing answers to quizzes integrated in the game [20], or analyze how players progress in learning-language courses to create rich visualizations for teachers [21].

We consider that serious game designers must take into account analytics and assessment constraints from the inception of the game and throughout the design phase [15]. Many authors have defined methodologies and guides to design serious games [13, 22–25], however, this body of research proposes methodologies that are applicable to any analytics-aware video game, serious or not. In particular, these works usually do not address key serious-game aspects, such as how to deliver knowledge and educational contents through gameplay or how to infer the corresponding learning outcomes. Some work is starting to explore these issues, proposing a taxonomy of possible elements that a serious game should include to be more effective [26].

To summarize, we found research that describes effective analytics-aware serious game design, but lacks concrete methodologies to infer learning outcomes. On the other hand, there is research that proposes ways to analyze serious game learning outcomes, either via general frameworks or ad-hoc

analysis, but without addressing the implications of that assessment in the game design. We propose to combine both approaches to define a methodology that tackles all the phases in the development of a serious game, from game design and implementation, to deployment and learning outcome analysis.

3. Proposed methodology

Our methodology pursues two goals: 1) to ease the measurement of serious game learning outcomes and 2) to provide a systematic way to assess the effectiveness of serious games as a whole. To achieve these goals, our approach covers the complete lifecycle of the serious game (Figure 1). The process starts in the design phase, where the learning goals and the target population are the basis to create a learning and game design. These designs combined are used to implement the game, which is then validated in a formative evaluation with a sample of the target population. This process is repeated until the game is fully validated. Then, the game is ready to be used by the target population (deployment). In the following subsections, we describe each step of the process in more detail.

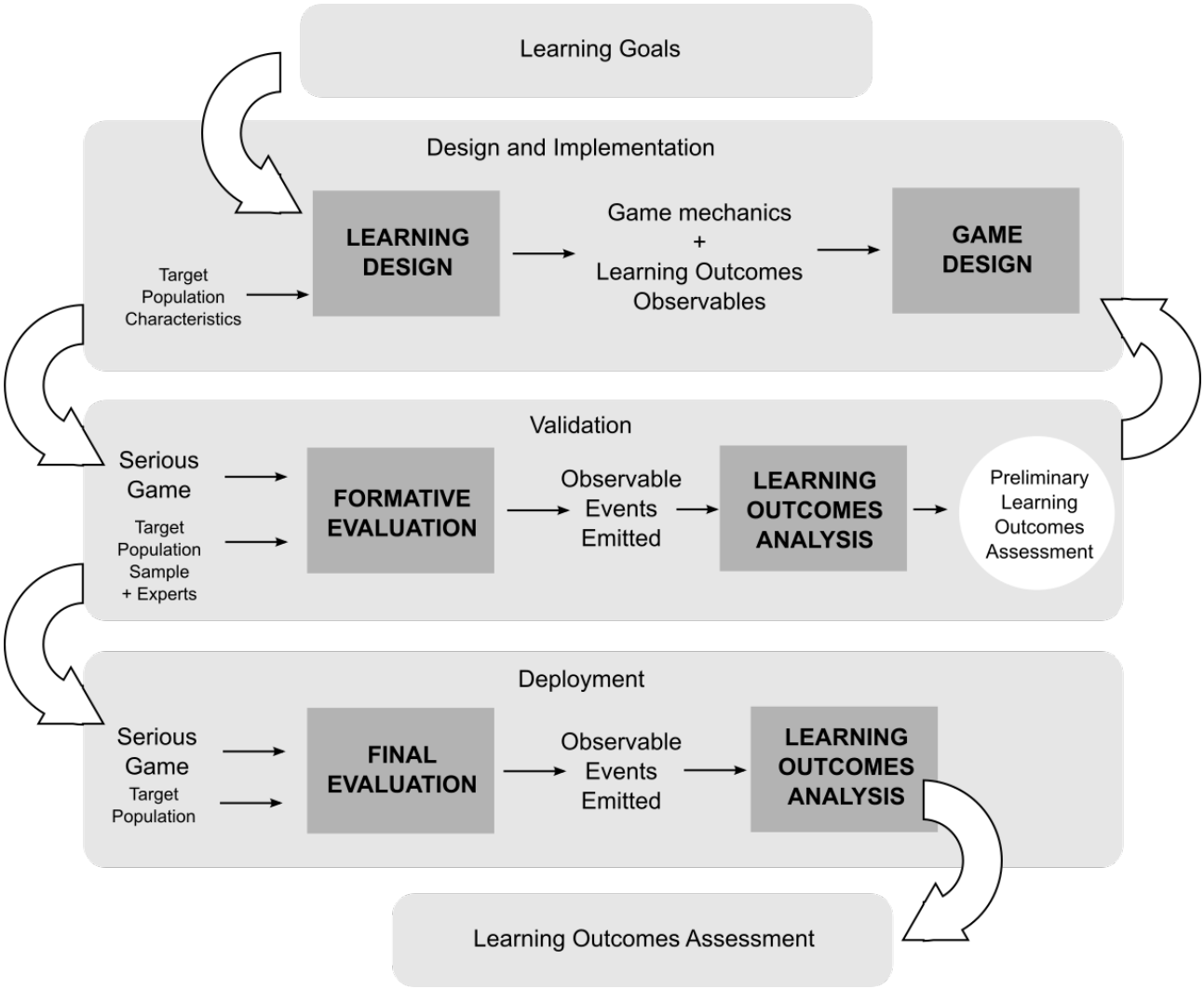

Figure 1. Serious game design and deployment process, with learning outcomes assessment.

3.1. Design and implementation

In the context of our methodology, we define “learning design” as the transformation of learning

goals into game mechanics and learning outcomes observables, considering the characteristics of the target population.

The chosen game mechanic should fulfill two requirements: 1) that is appropriate for the learning goal content, using models like the one presented in [27], where learning mechanics are mapped to game mechanics; and 2) that players' gameplay can produce learning outcome observables (also termed *events*) that attest the players' knowledge or skill.

During game design, these constraints, along with many other considerations for the game (such as art style, storytelling or technologies), shape the implementation of the serious game. Additionally, in this phase designers must define how the serious game should scaffold the delivery of the learning goals. Although there is a lack of concrete methodologies to translate educational theories into game design aspects [28], some authors have proposed models describing learning processes in

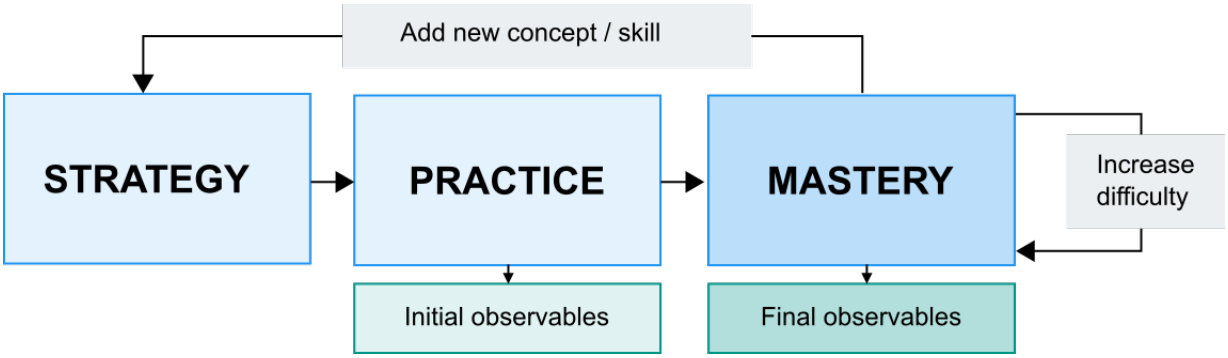

Figure 2. Game design pattern phases. The phases would be repeated for each sub-goal. Observables are emitted during practice and mastery phases.

videogames. For instance, in the serious games domain, Kiili proposes the experiential gaming model [29] and problem-based learning [28], both of them based on an iterative process in which players form an strategy, experiment in the game world, receive feedback, and reflect on the results. In the commercial videogames domain, there are similar proposals that split the experimentation in two sub-steps: experimentation in a "safe game environment", where the level of the difficulty of the challenge to overcome is low and mistakes are not punished, and experimentation in an "unsafe game environment", where the level of the difficulty is higher, and mistakes are punished (e.g., losing game lives, coins, score, etc.) [30].

For the purposes of our methodology, we have combined and extended these ideas into a game design pattern that also considers the learning outcomes observables. Each learning goal is presented to players throughout 3 phases, with 2 points of non-disruptive measurement (Figure 2):

1) **Strategy**: Players are first introduced to the learning goal. This can include knowledge they might need in subsequent steps, as well as concrete instructions on how to interact with the game world, for instance through non-interactive scenes or game tutorials. The player receives information to understand the challenge behind the learning goal and start forming some strategies to tackle it. This is also coherent with the initial exploration behavior that is very common in games.

2) **Practice**: players start to apply the knowledge presented in the previous phase. This practice must occur in a game environment where players' mistakes have either no consequence at all

or only mildly adverse consequences (“safe game environment”). This experimentation must be designed in such a way that players can make deductions and test hypothesis on both the knowledge presented in the previous phase and the game mechanics. In this phase students test and practice their strategies. Strategies that work better will later be refined by the player during the mastery phase.

In this phase, players apply the knowledge associated with the current learning goal for the first time. This allows us to collect initial observables from which their initial knowledge can be estimated.

3) **Mastery**: players are required to prove that they have acquired the intended knowledge, facing challenges similar to those presented in the practice phase, but with increasing difficulty, and greater in-game consequences, such as loss of score or in-game “lives” (“unsafe game environment”).

In this phase, players prove the degree to which they have acquired the intended skill or knowledge – therefore, we can collect final observables that will allow us to measure their final progress towards the learning goal.

These three phases can be iterated to deliver multiple learning goals, or to deliver a single goal with increasing difficulty, adding a new related concept or skill in each cycle. Additionally, this game pattern optimizes the time the players are in the flow zone [9], since it alternates moments where players are learning new things in a safe environment (practice), with moments where they are challenged to prove their skills (mastery), all with an incremental approach to avoid frustration.

3.2. Collecting observables

Players perform different interactions to advance in the game: they make choices, resolve puzzles, beat bosses, etc. These events will be the core observables to perform the learning outcomes analysis. The following principles (many of them shared with general GA) can facilitate this analysis:

1. Observable’s data should be time-stamped events, representing simple interactions of the player with the game [18]. These events should be sent to a central server, where all player interactions will be stored for later access and analysis.
2. Events sent to the server should be raw interactions instead of opaque scores [18, 31]. For instance, if the mastery phase contains two puzzles, the events to transmit would be the interactions performed to resolve the two puzzles, instead of just a combined score of the final result. This ensures flexibility, since scores can be later recalculated from interaction data if the subsequent analysis identifies a need to do so.
3. Data collection should be as non-disruptive during gameplay as possible. Ideally, game flow should never be interrupted to collect data – players should not be explicitly asked to stop their play to pass an exam or to answer questions not integrated in the gameplay.

Once all interaction events (observables) are stored in a central location, analysis can begin.

3.3. Learning outcome analysis

We store all gameplay interactions in a single server. Following our design pattern, each interaction is associated with a learning phase (strategy, practice or mastery) of a specific learning goal. Interactions from the strategy phase are not used to infer learning outcomes (since this phase should only contextualize the learning goal). Analyzing interactions from the other two phases, we can calculate two assessment scores:

1. **Initial assessment** (*IA*), using the initial observables from the practice phase. It estimates the learner's initial degree of knowledge. A high value would indicate that the player likely possessed this knowledge before starting to play, while a low value would mean the opposite.
2. **Final assessment** (*FA*) using the final observables from the mastery phase. It estimates the learning outcome. A high value would indicate the player succeeded in the learning goal, while a low value would indicate she failed.

The specific steps to transform observables events into *IA* and *FA* will be different for each serious game. However, they can generally be expressed with formulas that combine data from each interaction. In section 4, we provide details of this process in a real case study.

We define two assessment thresholds: an initial threshold (*IT*) associated with the *IA*, and a final threshold (*FT*) associated with the *FA*. These thresholds are used to determine whether a phase is successfully accomplished or not. For instance, if *FA*'s value ranges from 0 to 1, a possible value for *FT* could be 0.5, so we consider that a player that reaches an *FA* value equal or greater than 0.5 have successfully completed the mastery phase.

For serious games that includes multiple learning goals, we can calculate their global *IA* and *FA* using a weighted average combining results from each learning goal: given a game with *N* educational goals, each with two assessments (IA_i, FA_i), two thresholds (IT_i, FT_i) and a weight (W_i), we can then calculate the global assessment value (*A*) for the initial and final assessments as:

$$A = \frac{\sum_{i=1}^{N} A_i \times W_i}{\sum_{i=1}^{N} W_i}$$

And the global threshold value (*T*) for initial and final thresholds as:

$$T = \frac{\sum_{i=1}^{N} T_i \times W_i}{\sum_{i=1}^{N} T_i}$$

With these values, we can now estimate learning outcomes and assess the serious games' effectiveness.

3.3.1. Inferring players' learning outcomes

The analysis of observables or signals provides two measures for each learning goal: *FA* and *IA*. With these values, we can calculate two concrete learning outcomes:

- ***FA* as the final score of the player**: We can use *FA* as a score or mark for the players when considered as students – essentially scoring what they know after playing the game. We should avoid using *IA* to calculate this mark. Although it represents players' knowledge,

using it to calculate final marks would be unfair, since *IA* takes into account mistakes committed during the practice phase, while a fair grade should only consider what students know at the end of the game, not what they ignored at the beginning.

- **The difference between accomplishments in the practice and mastery phase as game effectiveness**: If we compare *IA* and *FA* to their respective thresholds (*IT* and *FT*), we can determine whether a player succeeded in the practice and mastery phase. The game is most effective when players that failed in the practice phase ended up succeeding in the mastery phase, as this indicates a knowledge gain. This difference is the base from which we calculate the serious game effectiveness.

3.3.2. Assessing serious game effectiveness

In the context of our methodology, we consider that a serious game is effective if we find a positive change in the knowledge level of the player. We can determine this change using the results of *IA* and *FA* respect to *IT* and *FT*. Using these values we can classify each player in a different learning category:

- If $FA \geq FT$, the players successfully completed the mastery phase and possess the skill intended for the learning goal. Depending on the *IA* value, we can classify players as either:
    - **Learners**, if $IA < IT$: players committed errors during the practice phase, indicating that they did not possess the skill or knowledge before playing the game. However, they ended up being successful in the master phase, which means there is an educational gain during the gameplay.
    - **Masters,** if $IA \geq IT$: the players did not commit errors during the practice phase, indicating that they probably possessed the skill or knowledge before playing.
- If $FA < FT$, the players failed the mastery phase and do not possess the skill intended for the educational goal. Depending on the *IA* value, we can classify players into two different categories:
    - **Non-Learners**, if $IA < IT$: the players also failed the practice phase, indicating that they struggled throughout the game potentially with little or no benefit.
    - **Outliers,** if $IA \geq IT$: the players succeeded during the practice phase, but were unable to apply the acquired knowledge in the mastery phase.

We determine serious game effectiveness by classifying each gameplay session according to these criteria, and then comparing the total number of players in each category.

If the majority of players were learners, the game was highly effective: most players learned something while playing. If the majority were masters, the game produced no learning effect since most players already possessed the intended knowledge before playing. If the majority were non-learners, the game was not effective at all, since most players were unable to success in any phase. And finally, a majority of outliers indicates that the game and/or the chosen *FA* and *IA* formulas probably had design flaws.

It is important to remark that most serious games will output different results for different

populations. A serious game could be highly effective with kids from 10 to 12, and not effective at all with kids from 7 to 9. The key is to have a well-defined target population during the design phase of the serious game, and to perform a validation process to ensure that effectiveness goals are met.

3.4. Validation and deployment

Once we have the serious game implemented along with the infrastructure to track its observables, hooked to the learning outcomes analysis, we need to validate it.

In the validation phase, domain experts and, ideally, a sample of the target population, play the serious game, producing gameplays that are later assessed with the learning outcomes analysis, yielding preliminary results, in a process usually called formative evaluation [32]. This process is iterative and designed to detect aspects to fix, polish, tweak or improve in the serious game, which can range from changing the game mechanic of a learning goal, because preliminary results suggest low performance, to changing the how *FA* and *IA* are calculated because experimental results contradict certain game design hypothesis.

Once the game is validated, it can be used in production in the final deployment. In this final phase, the serious game and its learning outcomes analysis are used to assess the students that play with it (final evaluation).

4. Case study

In previous sections, we have presented our methodology to model and infer learning outcomes and effectiveness in serious games. This section describes a case study that illustrates how this methodology works when applied. The case study starts with the following research questions:

RQ1.What are the implications of using our game-design pattern during the design and implementation of a serious game?

RQ2. What results, regarding learning outcomes and effectiveness, can be obtained from a serious game developed and analyzed with this methodology?

To answer them, we used the proposed methodology to implement and analyze “The Foolish Lady”, a serious game[1] based on the homonymous theater play by Spanish playwright Lope de Vega. In this game, players are presented with several language and literature challenges. Its main learning goal is to teach youngsters about Spanish Golden Century poetry. In the following subsections, we describe the design and implementation process, the data collection and analysis process, and the results of an experiment with 320 high school students that played the game.

4.1. Design and implementation

“The Foolish Lady” serious game [33, 34] is an adventure game based on a classical Spanish play. In the game, players advance through scenes of the play making decisions that affect the overall story and the final scene. Along the way, they find puzzles and mini-games where they need to apply knowledge on language and literature. The game is designed to be completed in 30 to 40 minutes.

[1] Available (in Spanish) at https://play.google.com/store/apps/details?id=es.eucm.androidgames.damaboba

One of its main learning goals is to teach poetry structure and rhymes, especially, the “redondilla”, a Spanish poetic composition that uses a specific rhyming scheme and verse length. During the learning design phase, the chosen game mechanic was point and click mini-games (drag-and-drop puzzles and option selection in conversations with non-playable characters in the game), typical of adventure games, due to the educational benefits of this genre [35]. During the game design, we subdivided this goal into the three phases defined by our design pattern. Figures 3, 4, and 5 show in-game screen captures for each of these phases.

Players are first presented with a written explanation with notions of rhymes and the “redondilla” structure (Figure 3). These instructions appear in two non-interactive scenes that can be skipped (after reading the content or not) with a click. These scenes belong to the *strategy phase*.

Later on, players find a mini-game where they need to complete a poem composed as a “redondilla” (Figure 4). The poem is missing five words and the players can fill in the blanks by dragging words from a container placed at the right side of the screen. Once filled, they can check the correctness of the poem by clicking a check button. They can try as many times as they want, until they find the right combination of words: as the *practice phase* of the goal, results in this mini-game are irrelevant for the final score.

Finally, players encounter two mini-games (Figure 5). In the first one, players must fight a knight by exchanging rhyming verses. Players win this battle if they choose three correct rhyming replies in a row – or lose it if they fail three times in a row. Their score decreases with each error. In the second and final mini-game, the foolish lady's father assesses the protagonist’s suitability as a son-in-law by asking the player a series of questions on the “redondilla” poetic composition. Players can answer these questions only once, and both the score and their protagonist’s marital prospects decrease if they fail. Both mini-games belong to the *mastery phase* of the goal, and therefore their results affect the final score.

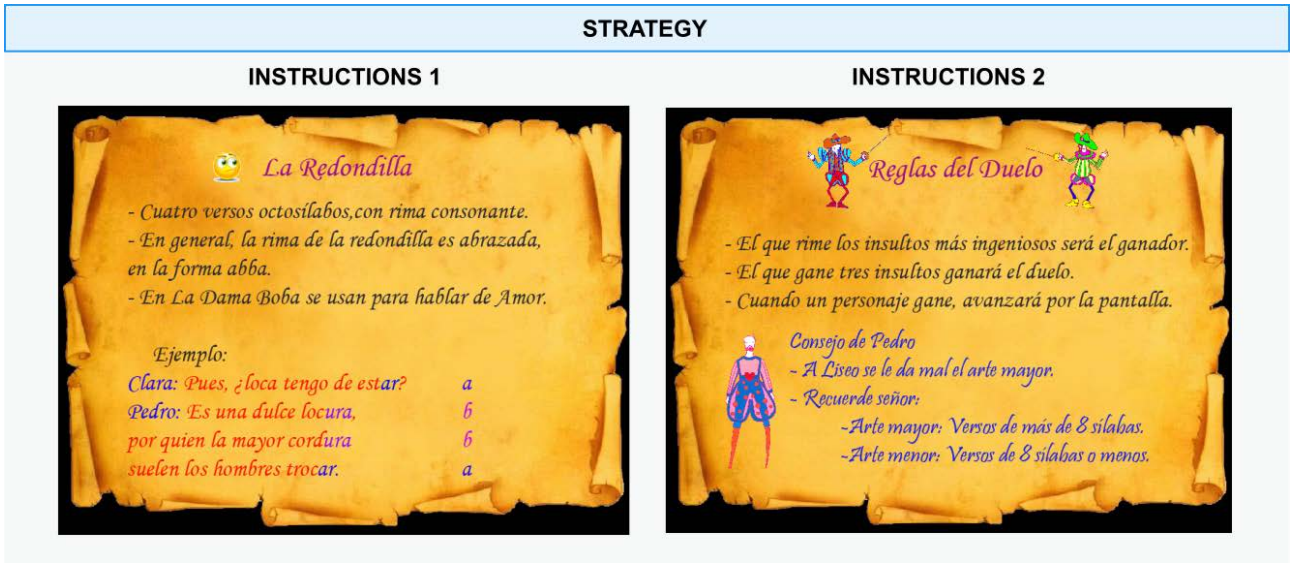

Figure 3. The game exposes the basic concepts of the “redondilla” through two screens with written explanations.

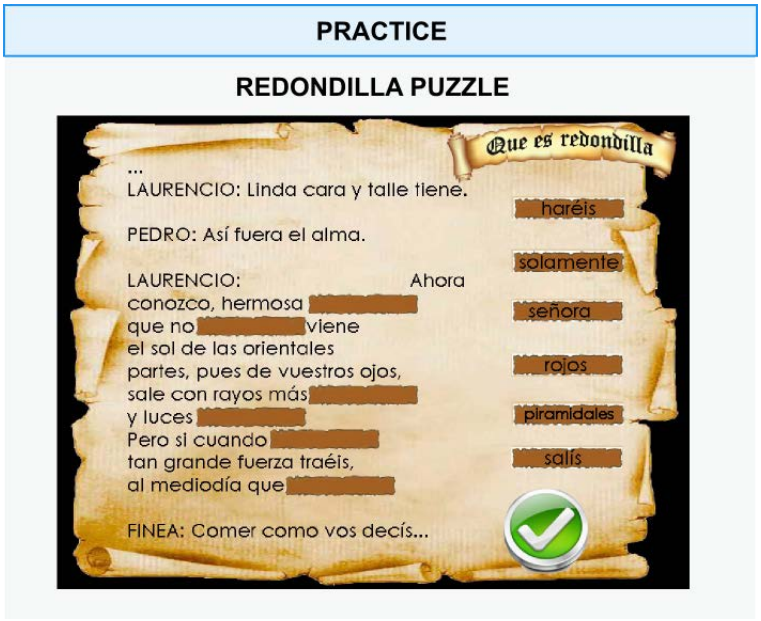

Figure 4. In the first puzzle, players need to apply their knowledge of the "redondilla". They can try as many times as they want.

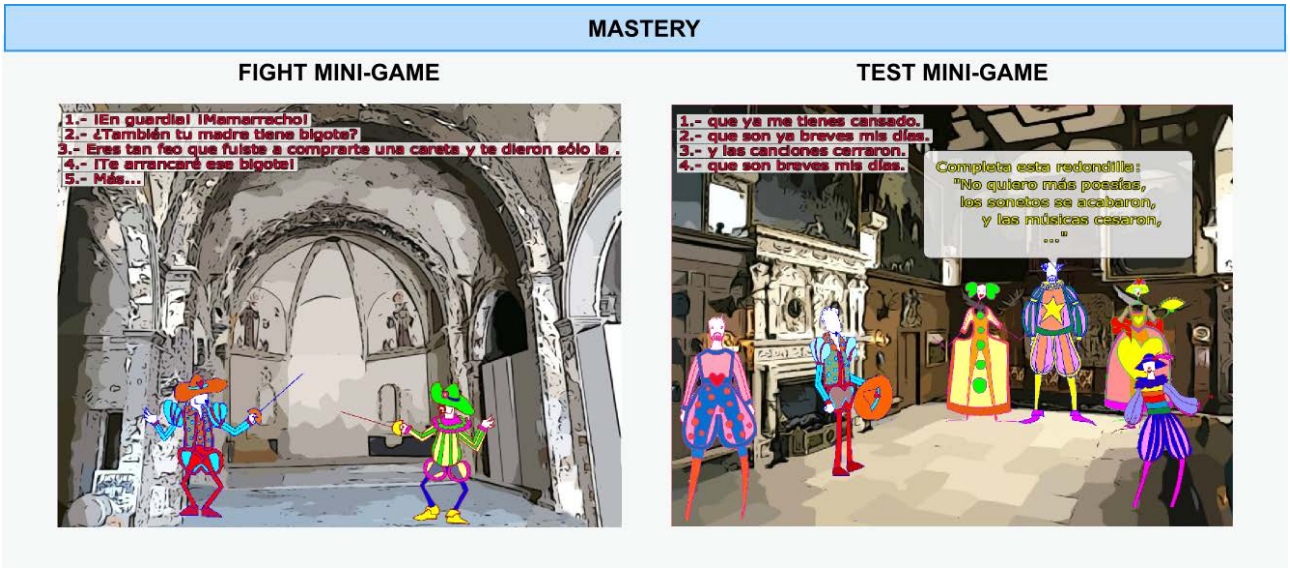

Figure 5. In the final mini-games, players must prove their knowledge. In both cases, players lose score with each error.

4.2. Collecting observables

To record and analyze the gameplay sessions of all students, we developed a framework composed by a *tracker,* bundled within the game itself responsible for sending events (observables), and by a *collector server,* responsible for receiving and storing the events. The type of events are fully detailed in [18, 31]; here, we only highlight those events relevant for the learning outcomes analysis:

- Events representing that a new attempt to beat the "redondilla puzzle" was started. Every time the player click the "check" button and the result is incorrect a new attempt starts.
- Events representing that a new attempt to beat the "fight mini-game" was started. Every time the player loses the fight and restarts the mini-game a new attempt starts.

- Answers chosen by the player during the final mini-game.

The game itself does not make any assessment calculation: only raw events are sent to the server.

4.3. Learning outcomes analysis

All players encounter the 3 mini-games during their playthroughs: the “redondilla” puzzle mini-game in the practice phase, and the fight mini-game and the test mini-game in the mastery phase. For each mini-game we calculate a score between 0 and 1:

- **Redondilla Game score (*RG*)**: if *A* is the observable representing the number of attempts to solve the “redondilla” puzzle mini-game, *RG* is computed using the formula $RG = 1 - (\text{MIN}(A - 1, A_{MAX}) / A_{MAX})$, where A_{MAX} is the reasonable number of attempts needed to solve the game. The initial assessment will be 1 when the player beats the puzzle at the first attempt, i.e., $A = 1$. The initial assessment will be 0 if the player does not complete any attempt on the puzzle, or tries over A_{MAX} times.
- **Fight Game score (*FG*)**: if *E* is the observable representing the number of erroneous options chosen before completing the fight mini-game, *FG* is calculated by the formula $FG = \text{MAX}(0, 1 - (\text{MIN}(E, E_{MAX}) / E_{MAX}))$, where E_{MAX} is the maximum number of reasonable errors needed to beat it.
- **Test Game score (*TG*):** In the test mini-game, each question has four answers, and only one of them is correct. Each answer has an associated score. The correct answer always has a score of 0, and the rest of answers have scores that correspond to their distance from the truth: 1 for answers that are almost right, 2 for answers that are wrong, and 3 for answers that, due to their content or formulation, are clearly intended as jokes. If *I* is the observable representing the accumulated score of incorrect answers after finishing the test mini-game, $TG = \text{MAX}(0, 1 – I / 4)$, since the number of questions asked is 4.

We set $A_{MAX} = 3$ and $E_{MAX} = 6$. These values were agreed between game designers and educators, considering the educational and game challenge each mini-game entails for the players. However, since we are going to track raw *A* and *E* values, A_{MAX} and E_{MAX} values can always be changed *a posteriori* if, after running the validation process, the data suggest that more appropriate values should be applied.

With these values, now we can calculate IA and FA:

- $IA = RG$, since the “redondilla” puzzle mini-game is the only one in the practice phase.
- $FA = FG \times 0.5 + TG \times 0.5$, since the fight and test mini-game are in the mastery phase, and we have decided to give both equal weights in the final score.

For all mini-games, we set the assessment threshold to 0.5, making both the *IA* and *FA* thresholds also 0.5.

Table 1 shows possible values for *RG*, *FG*, *TG*, *IA* and *FA* used in the analysis of this experiment.

Initial Assessment / Redondilla game $A_{MAX}=3$ $IA = RG= 1 – (MIN(A - 1, A_{MAX}) / A_{MAX})$		***Final Assessment (FA)***				
		Fight game $E_{MAX}=6$ $FG= 1 – (MIN(E, E_{MAX}) / E_{MAX})$		***Test game*** $TG = MAX(0, 1 - I/4)$		
Attempts (A)	***RG/IA***	*Errors (E)*	*FG*	*Incorrect score (I)*	*TG*	***FINAL ASSESS-MENT*** *0.5*FG+0.5*TG*
1	***1***	*0*	*1*	*0*	*1*	***1***
2	***.66***	*1*	*.83*	*1*	*.75*	***0.8***
3	***.33***	*2*	*.66*	*2*	*.5*	***0.58***
4	***0***	*3*	*.5*	*3*	*.25*	***0.375***
5	***0***	*4*	*0.33*	*4*	*0*	***0.165***
6	***0***	*5*	*0.16*	*5*	*0*	***0.08***
7 or more	***0***	*6 or more*	*0*	*6 or more*	*0*	***0***

Table 1: Some illustrative values for *IA* and the components of *FA*.

4.4. Case study

To answer RQ2, we ran an experiment with high school students that played the serious game.

4.4.1. Experimental design

Before high school students (our target population) played the game, and as part of the validation process, we first ran a formative evaluation with graduate students [34] and the teachers involved in the experiment. Results from this validation helped to fix some implementation flaws, and to improve the gameplay and the overall learning design. For instance, two questions from the final mini-game were changed to improve its alignment with the learning goal.

After the validation, high school students played “The Foolish Lady” during 30 to 40 minutes in a PC, under the supervision of a researcher who did not provide any sort of assistance (only a brief indication on how to start the game). We collected one gameplay per student (deployment phase). We consider a gameplay session as the set of traces (interactions with the game) generated from the first screen to the final screen of the game.

From each gameplay, we computed 3 values: *RG*, *FG* and *TG*. Those that did not complete a mini-game scored 0. From these variables, we calculated *IA* and *FA* with the formulas presented above. Using their results, we classified each student in a learning category (learner, master, non-learner or outlier) to draw conclusions on the game’s effectiveness.

In order to gain insight into our methodology, we wanted to know if we could answer these case study questions (CSQ) concerning “The Foolish Lady” serious game:

- **CS1**: **Did the students possess the intended skill at the end of “The Foolish Lady” game? Given our demographic variables, were there differences between groups?**
- **CS2: Is “The Foolish Lady” game effective at teaching its intended skill to our population? Given our demographic variables, were there differences between groups?**

4.4.2. Participants

The experiment involved N = 320 high school students from 8 different schools in Madrid. 32 of the gameplay sessions were corrupted or incomplete, due to different technical problems while playing the game (power cuts, Internet connection issues and computer malfunctions), and were therefore discarded.

The gender proportion in the resulting population (N = 288) was 44.4% females and 55.6% males. The participants were all between the ages of 12 and 16 (mean age was 13.70 ± 1.27), from high schools of the Madrid area. By schools, 3 were charter or private schools (58% of the population), and 4 were public schools (42% of the population). In terms of gender, age and school type, the participants are a representative sample of the student population of Madrid for this age [36, 37].

Additionally, we collected participant game habits to classify each student in a player category, evaluating what types of games they play and how often. According to the instrument developed by [38], 14.9% were non-gamers (they never play any type of video games), 28.8% were casual gamers (they play casual video games for short periods of times), 31.6 % were hardcore gamers (they frequently play games such as FPS or MMORPG) and 24.6% were well-rounded gamers (they play all types of games frequently). There is a more detailed explanation of each category in [38].

5. Results

In this section, we present the results of the learning outcomes analysis of the deployment phase, i.e., results from the high school students.

5.1. Game completion

Figure 6 shows the number of players that completed each phase of "The Foolish Lady": all 288 players started the game and also completed the strategy phase; 281 completed the "redondilla" puzzle mini-game; 246 completed the fight mini-game; and 231 completed the test mini-game. The largest drop of players (35) happens between the "redondilla" puzzle and the fight mini-game.

In summary, 80.21% of players finished the game at least once.

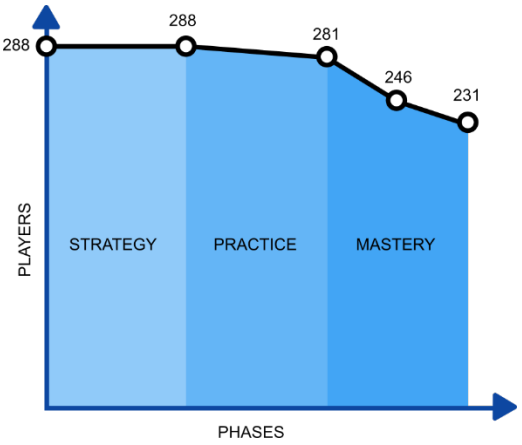

Figure 6: Number of players that accomplished each phase of "The Foolish Lady" game.

5.2. Learning outcomes

To answer whether the students reached the intended skill level at the end of the game, we calculated the values of *RG*, *FG* and *TG*, and therefore, *FA* and *IA*. In total, 196 players (68.05% of the total population, 84.84% of players that completed the game) scored more than 0.5 (adequacy threshold set for the game during design) in *FA*. *IA*'s mean value is greater than 0.5 in all age groups.

On the other hand, the second part of CS1 led us to calculate *FA* and *IA* across the different demographic groups: gender, age and gaming-habits. We first performed a one-way analysis of

variance (ANOVA) over *IA*, to find initial differences across groups. *IA* showed statistically significant differences on groups by gender and game habits (Table 2).Therefore, to consider these differences on *IA*, we applied an analysis of covariance (ANCOVA) to evaluate differences on the *FA* score (dependent variable), across groups (independent variables) using the *IA* score as covariate. Standard preliminary checks were conducted to confirm that there was no violation of the assumptions of normality, linearity, homogeneity of variances and homogeneity of regression slopes [39].

Table 3 shows the ANCOVA results for the 3 independent variables, showing statistically significant differences ($p<0.05$) among groups by age and game habits: First ANCOVA [between-subjects factor: age (12 to 16); covariate: IA scores] revealed main effects of age $F(4,288) =7.28$, $p<0.01$, and a medium $\eta_p^2=.094$; second ANCOVA [between-subjects factor: gender (male, female); covariate: IA scores] showed no main effects on gender $F (1,288) =.62$, $p=.43$, $\eta_p^2=.002$; and third ANCOVA [between-subjects factor: game-habits (4 clusters); covariate: IA scores] revealed main effects of game-habits $F(3, 288) =2.880$, $p=.036$, and a small $\eta_p^2= .030$. Table 4 shows the adjusted means for each demographic group.

	One-way ANOVAs on IA			
Independent variable	*N*	*df*	*F*	*p*
Age	288	4	2.5	.031
Gender	288	1	18.41	<.005
Game Habits	288	3	12.10	<.005

Table 2: ANOVA results on *IA*, showing significant differences among groups by gender and game habits.

	ANCOVAs on FA				
Independent variable	*N*	*df*	*F*	*p*	*Partial η²*
Age	288	4	7.28	.000*	.094
Gender	288	1	.62	.43	.002
Game Habits	288	3	2.88	.036*	.030

*p<0.05

Table 3: Test scores and ANCOVA results by age, gender and gaming profile.

		ANCOVA		
Ind. Variable	*Values*	*N*	*Adj. Mean**	*Std. Err.*
Age	12	69	.508	.038
	13	50	.508	.044
	14	96	.708	.032
	15	43	.631	.048
	16	30	.766	.057
Gender	Female	128	.603	.029
	Male	160	.633	.026
Game Habits	Casual	83	.559	.035
	Non-gamer	43	.554	.049
	Well-rounded gamer	71	.680	.038
	Hardcore	91	.659	.034

*Adjusted mean using practice phase scores as covariate (ia=.6146)

Table 4: *FA* adjusted means by age, gender and gaming profile.

5.3. Serious game effectiveness

Figure 7 shows the total number of players grouped by learning category. Most players are masters, followed by learners. The number of outliers is higher than that of non-learners.

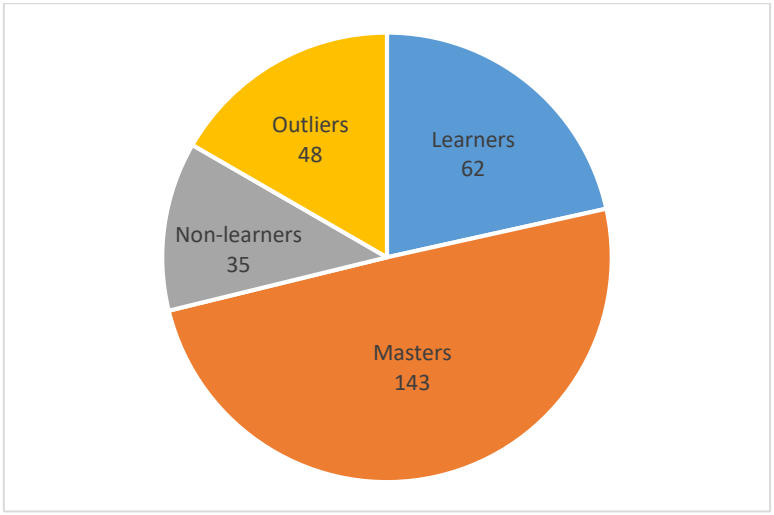

Figure 7. Categorization of players according to their assessment category.

Figure 8 and 9 shows the players' categorization grouped by learning category and segmented by age and game-habits. In all groups, the number of masters exceeds that of other categories, especially in the 14 year-old group. In all groups, the number of outliers is greater than the number of non-learners, except in the group of students aged 16.

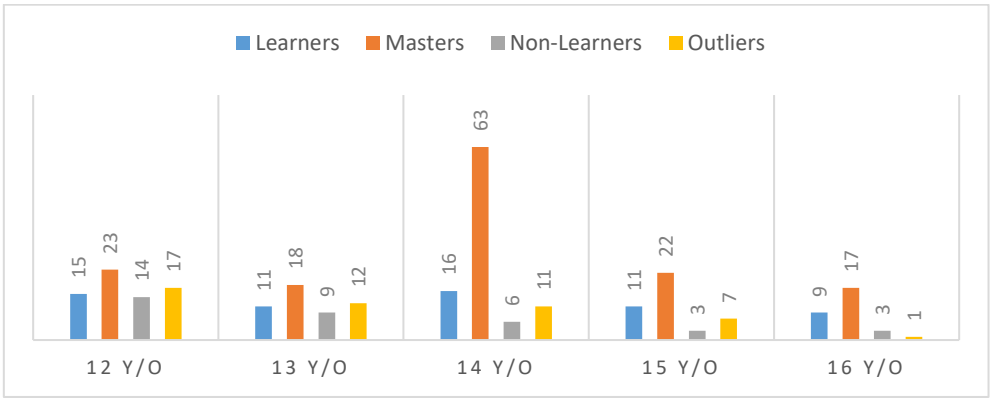

Figure 8. Distribution of players across assessment categories, segmented by age.

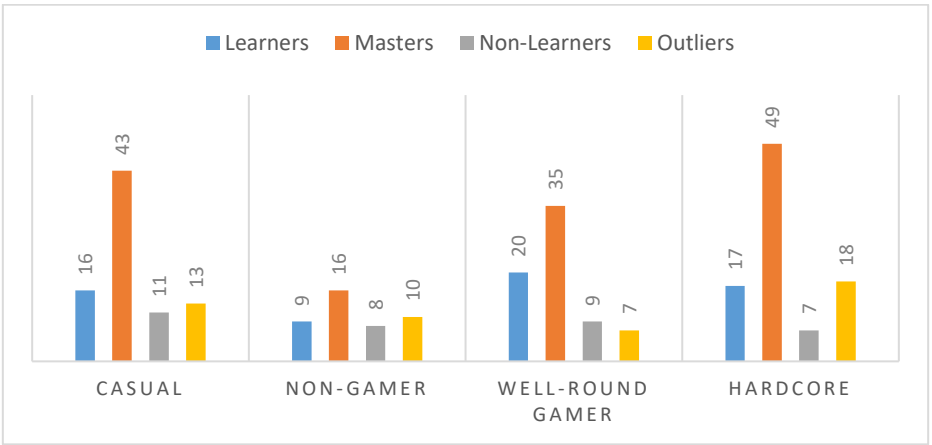

Figure 9. Distribution of players across assessment categories, segmented by game-habits.

6. Discussion

In this section we first present the answers to the case study questions, and then further elaborate to answer the methodology research questions.

CS1: Did the students possess the intended skill at the end of "The Foolish Lady" game? Given our demographic variables, were there differences between groups?

Yes. 80.2% of students completed the game, which required, by design, a basic understanding of the principles of the learning goal. Regarding the final score of these students, ANCOVA analysis (dep. Variable: FA; covariate: IA) has shown statistical differences when data are segmented by age and by gaming habits.

By age, students aged 12 and 13 obtained the lowest values (Adj. Mean= .508), and students aged 16 obtained the highest values (Adj. Mean=.766). It seems natural: older students found the game easier.

By game-habits, the segment with the best results are well-round gamers (Adj. Mean= .680), closely followed by hardcore gamers (Adj. Mean= .659). These two types of players are used to playing games with complex mechanics. "The Foolish Lady" is an adventure game, with fairly simple mechanics, so these players' expertise probably helped them to complete the game more effectively.

CS2: Is "The Foolish Lady" game effective at teaching its intended skill to our population? Given our demographic variables, were there differences between groups?

No. Not because the players did not learn, but because, according to the results, most of them were categorized as "Masters", i.e., many of them already knew most of the educational contents. This could mean the game was too easy for most of the players. However, we think there is an additional problem in the game design that prevented us from capturing a more accurate IA (and, consequently, a more accurate learning profile): since we wanted to keep the game short —so that it could be completed in 40-minute sessions—the practice phase was deliberately shorter than the mastery phase. This forced us to keep it only to a single, comparatively easy mini-game, whose score was not enough to fully measure the initial knowledge. This flaw went unnoticed during the validation process because the players in this phase were domain experts, and therefore, were classified as masters, which seemed natural. That is why the serious game should be also validated with a sample of the target population.

Segmenting groups by age and game-habits there is no particular group in which the game was more effective.

These results do not imply the game has no value as educational tool. Students playing this game enjoyed other benefits, such as a measurable increase their motivation to go to the theatre, as demonstrated in [33].

RQ1. What are the implications of using our game-design pattern during the design and implementation of a serious game?

The methodology forced us to define a clear learning goal from the beginning, and to stick to it during the game development process.

In cooperation with the educational experts, we designed the mini-games clearly defining which role each of them covered in delivering the educational goal. We defined mini-game difficulty, weight and placement guided by the proposed game-design pattern. The mini-games were implemented in such a way the interactions and events involved in their resolution were clearly identified, and since the beginning, we formulated how those events were converted into assessment.

We also integrated a tracker into the game engine to capture all relevant interactions. This approach is common in the games industry for any analytics-related tasks, although its difficulty varies depending on the chosen game engine. In our case, we used an open source engine, where all the required events and interactions were generated in a handful of locations within the code; wiring in the tracker was relatively simple.

We also needed a service to collect all the data. Ours consisted of a REST back-end to process HTTP requests with the events, a database to store these traces, and some Python scripts to query the database. Although we used a custom solution, the serious game could be integrated in any other VLE. This opens new interesting questions regarding sharing of data between these systems and the serious game, which however fall out of the scope of this paper.

RQ2. What results, regarding learning outcomes and effectiveness, can be obtained from a serious game developed and analyzed with our methodology?

Identifying relevant educational observables during game design simplified the task of calculating learning outcomes and game effectiveness. These results were used to answer several interesting questions of the case study.

By default, our methodology converts the serious game into an assessment tool: it relies on clear assessment locations that are associated with both the learning design and goals, which are then combined to infer learning outcomes. However, using our design pattern, we can also determine if students actually learned playing the game, which is key to assess the game's effectiveness within a particular population group.

In our case study, we concluded that the initial assessment was higher than expected. The number of outliers indicates a design flaw in the practice phase. We consider that this finding based on actual data is very useful. Once this flaw is detected, we can iterate over the methodology again to improve the game.

Finally, if student demographics data is available, we can use statistical analysis to identify and characterize those groups in which the learning outcomes are better or where they game effect is higher. This helps to narrow down and better identify the ideal target population for the game.

7. Conclusions

In this paper we present a methodology to structure the design and assessment of serious games at two levels: inferring learning outcomes and assessing serious games effectiveness as educational tools. We think this is a contribution to systematize serious games development that improves some

of the aspects of the methodologies found in the literature review: our methodology is fully integrated with the production cycle of a serious game (from design to deployment), and proposes a non-disruptive assessment alternative to questionnaires, the most common assessment method for serious games. It poses extra requirements during game development (a tracker in the game engine and a server to collect the data), but with today's big data technologies this becomes an affordable task.

We tested our methodology developing a serious game that was played by 320 students. The methodology clearly guided our steps in the design process, and later in the analysis that we would use to determine if "The Foolish Lady" was an effective serious game. The game proved to be an effective assessment tool (i.e., we were able to give a mark to each student), however, it was unable to fully capture the initial knowledge of the students.

One of the conclusions of the experiment is that the design of the practice phase is key to implement an effective serious game. However, the balance in the practice phase can be hard to keep: the designer wants the player to advance flawlessly, while capturing their mistakes to obtain an accurate assessment of initial knowledge. Additionally, the implementation result derived from the case study (i.e., a tracker and a basic server infrastructure to receive and analyze traces) is going to be used in the RAGE European Project [40] as main infrastructure to assess games.

Although our case study is focused on a serious games designed to deliver knowledge and teach several skills, we think the methodology could be applied to any serious game whose goal can be measured in a quantitative way. For instance, a serious game designed to help diabetics to control their blood glucose levels could ask for the players' levels to determine whether the goal was achieved (instead of relying on puzzle or mini-games, as our game does).

In summary, we consider that the methodology presented in this paper provides a richer and a more understandable assessment analysis for serious games. One major point is that, once the game starts sending observable events, everything is automated, and all assessments are based on how learners interact with the game, instead of using traditional out-of-game questionnaires. Additionally, the assessment model is adaptable to researchers' needs since it is not hardwired to game signals: the way each dependent variable (*FA* and *IA*) is calculated can be changed a posteriori, allowing the constants used in assessment model to be updated if required (for instance A_{MAX} and E_{MAX}, in our case study). Additionally, results obtained by this methodology could complement formal experiments to measure serious games effectiveness, which is still an open issue [41].

We believe that this methodology opens up new research venues. In this paper we have limited the students' assessments to 3 particular points for clarity reasons (the 3 mini-games). In the future, we plan to enrich our game design pattern with more observables in of both phases. These data will provide us with more information on students' progression, enabling researchers to build a more precise photo of what is exactly going on during the learning process. We plan to go one step further by analyzing other gameplay data (such as the time spent in each phase), that may shed light into the reasons that make some players struggle in certain areas of the game. We also want to explore further the transformation from game observables to assessment scores, by identifying and addressing common patterns in different game mechanics.

Finally, the integration of serious games following our proposed methodology inside VLE also

raises interesting questions. What standards should be used in the communication? What visualizations should be provided to the different stakeholders? Addressing this integration will be an important step towards realizing the full potential of combining serious games and learning analytics.

8. References

1. Liu G, Fu L, Rode A V., Craig VSJ (2011) Water Droplet Motion Control on Superhydrophobic Surfaces: Exploiting the Wenzel-to-Cassie Transition. Langmuir. doi: 10.1021/la104669k
2. Squire K (2003) Video games in education. International Journal of Intelligent. Simulations and Gaming 2:49–62.
3. Connolly TM, Boyle E a., MacArthur E, et al (2012) A systematic literature review of empirical evidence on computer games and serious games. Comput Educ 59:661–686. doi: 10.1016/j.compedu.2012.03.004
4. Loh CS, Sheng Y, Ifenthaler D (2015) Serious Games Analytics: Theoretical Framework. In: Serious Games Anal. Springer International Publishing, Cham, pp 3–29
5. Elias T (2011) Learning Analytics : Definitions , Processes and Potential. Learning 23. doi: 10.1.1.456.7092
6. Chatti MA, Dyckhoff AL, Schroeder U, Thüs H (2012) A reference model for learning analytics. Int J Technol Enhanc Learn 4:318–331. doi: 10.1504/IJTEL.2012.051815
7. Ferguson R (2012) The state of learning analytics in 2012: a review and future challenges. Tech Rep KMI-12-01. doi: 10.1504/IJTEL.2012.051816
8. El-Nasr MS, Drachen A, Canossa A (2013) Game Analytics: Maximizing the Value of Player Data. doi: 10.1007/978-1-4471-4769-5
9. Chen J (2007) Flow in games (and everything else). Commun ACM 50:31. doi: 10.1145/1232743.1232769
10. Santhosh S, Vaden M (2013) Telemetry and Analytics Best Practices and Lessons Learned. In: Game Anal. Maximizing Value Play. Data. pp 85–109
11. Calderón A, Ruiz M (2015) A systematic literature review on serious games evaluation: An application to software project management. Comput Educ 87:396–422. doi: 10.1016/j.compedu.2015.07.011
12. All A, Nuñez Castellar EP, Van Looy J (2015) Towards a conceptual framework for assessing the effectiveness of digital game-based learning. Comput Educ 88:29–37. doi: 10.1016/j.compedu.2015.04.012
13. Annetta LA (2010) The “I’s” have it: A framework for serious educational game design. Rev Gen Psychol 14:105–112. doi: 10.1037/a0018985
14. Vargas JA, García-Mundo L, Genero M, Piattini M (2014) A Systematic Mapping Study on Serious Game Quality. In: Proc. 18th Int. Conf. Eval. Assess. Softw. Eng. EASE 2014. pp 1–10
15. Moreno-Ger P, Burgos D, Martínez-Ortiz I, et al (2008) Educational game design for online education. Comput Human Behav 24:2530–2540. doi: 10.1016/j.chb.2008.03.012
16. Hauge JB, Berta R, Fiucci G, et al (2014) Implications of Learning Analytics for Serious Game Design. In: Proc. 14th Int. Conf. Adv. Learn. Technol. IEEE, pp 230–232
17. Owen VE, Ramirez D, Salmon A, Halverson R (2014) Capturing Learner Trajectories in Educational Games through ADAGE (Assessment Data Aggregator for Game

Environments): A Click-Stream Data Framework for Assessment of Learning in Play. Am Educ Res Assoc Annu Meet 1–7.

18. Serrano Á, Marchiori EJ, Blanco Á del, et al (2012) A framework to improve evaluation in educational games. In: IEEE Glob. Eng. Educ. Conf. IEEE, pp 1–8
19. Lee SJ, Liu Y, Popovic Z (2014) Learning Individual Behavior in an Educational Game : A Data-Driven Approach. In: Proc. 7th Int. Conf. Educ. Data Min. pp 114–121
20. Dudzinski M, Greenhill D, Kayyali R, et al (2013) The Design and Evaluation of a Multiplayer Serious Game for Pharmacy Students. In: Proc. 7th Eur. Conf. Games Based Learn. Vols 1 2. pp 140–148
21. Ye F (2014) Validity, reliability, and concordance of the Duolingo English Test. https://s3.amazonaws.com/duolingo-certifications-data/CorrelationStudy.pdf. Accessed 27 Nov 2016
22. Marne B, Wisdom J, Huynh-Kim-Bang B, Labat J-M (2012) The six facets of serious game design: a methodology enhanced by our design pattern library. In: 21st Century Learn. 21st Century Ski. pp 208–221
23. Dickey MD (2006) Game design and learning: a conjectural analysis of how massively multiple online role-playing games (MMORPGs) foster intrinsic motivation. Educ Technol Res Dev 55:253–273. doi: 10.1007/s11423-006-9004-7
24. Denis G, Jouvelot P (2005) Motivation-driven educational game design. In: Proc. 2005 ACM SIGCHI Int. Conf. Adv. Comput. Entertain. Technol. - ACE '05. ACM Press, New York, New York, USA, pp 462–465
25. Dondlinger M (2007) Educational video game design: A review of the literature. J Appl Educ Technol 4:21–31. doi: 10.1108/10748120410540463
26. Carvalho MB, Bellotti F, Berta R, et al (2015) An activity theory-based model for serious games analysis and conceptual design. Comput Educ 87:166–181. doi: 10.1016/j.compedu.2015.03.023
27. Arnab S, Lim T, Carvalho MB, et al (2015) Mapping learning and game mechanics for serious games analysis. Br J Educ Technol 46:391–411. doi: 10.1111/bjet.12113
28. Kiili K, Ketamo H (2007) Exploring the learning mechanism in educational games. In: Proc. Int. Conf. Inf. Technol. Interfaces, ITI. pp 357–362
29. Kiili K (2005) Digital game-based learning: Towards an experiential gaming model. Internet High Educ 8:13–24. doi: 10.1016/j.iheduc.2004.12.001
30. Nutt C, Hayashida K (2012) The Structure of Fun: Learning from Super Mario 3D Land's Director. In: Gamasutra. http://www.gamasutra.com/view/feature/168460/the_structure_of_fun_learning_.php?page=4. Accessed 27 Nov 2016
31. Serrano-Laguna Á, Torrente J, Moreno-Ger P, Manjón BF (2012) Tracing a little for big improvements: Application of learning analytics and videogames for student assessment. In: Procedia Comput. Sci. pp 203–209
32. Fuchs LS, Fuchs D (1986) Effects of Systematic Formative Evaluation: a Meta-Analysis. Except Child 53:199–208. doi: 10.1177/001440298605300301
33. Manero B, Torrente J, Serrano Á, et al (2015) Can educational video games increase high school students' interest in theatre? Comput Educ 87:182–191. doi: http://dx.doi.org/10.1016/j.compedu.2015.06.006
34. Manero B, Fernández-Vara C, Fernández-Manjón B (2013) E-learning a escena: De La Dama Boba a Juego Serio. Vaep Rita 1:51–58.

35. Dickey MD (2006) Game design narrative for learning: Appropriating adventure game design narrative devices and techniques for the design of interactive learning environments. Educ Technol Res Dev 54:245–263. doi: 10.1007/s11423-006-8806-y

36. Comunidad de Madrid (2011) Datos y Cifras de la Educación. http://www.madrid.org/cs/Satellite?blobcol=urldata&blobheader=application/pdf&blobheadername1=Content-Disposition&blobheadervalue1=filename=DATOS+Y+CIFRAS+2010_2011.pdf&blobkey=id&blobtable=MungoBlobs&blobwhere=1271936872331&ssbinary=true. Accessed 27 Nov 2016

37. Ministerio de Educación (2008) Escolarización y población. http://www.mecd.gob.es/dctm/ievaluacion/indicadores/2011-e1.2.pdf?documentId=0901e72b810b4d41. Accessed 27 Nov 2016

38. Manero B, Torrente J, Fernández-Vara C, Fernández-Manjón B (2015) Gaming preferences and habits, gender and age on educational videogames effectiveness: An exploratory study (In press). IEEE Trans. Learn. Technol.

39. Pallant J (2013) SPSS survival manual: a step by step guide to data analysis using IBM SPSS. Open Univ Pr

40. Hollins P, Westera W, Manero B (2015) Amplifying applied game development and uptake.

41. All A, Nuñez Castellar EP, Van Looy J (2016) Assessing the effectiveness of digital game-based learning: Best practices. Comput Educ 92–93:90–103. doi: 10.1016/j.compedu.2015.10.007