

A methodology for assessing the effectiveness of serious games and for inferring player learning outcomes

Ángel Serrano-Laguna¹, Borja Manero¹, Manuel Freire¹, Baltasar Fernández-Manjón¹

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 to serve as educational tools in many educational domains, there is a lack of reliable, automated and repeatable methodologies that measure their effectiveness: what do players know after playing serious games? Do they learn from them? Previous research shows that the vast majority of serious games are assessed by using questionnaires, which is in stark contrast to current trends in the video game industry. Commercial videogame developers have been learning from their players through Game Analytics for years via non-disruptive game tracking. In this paper, we propose a methodology for assessing serious game effectiveness based on non-disruptive in-game tracking. The methodology involves a design pattern that structures the delivery of educational goals through a game. This structure also allows one to infer learning outcomes for each individual player, which, when aggregated, determine the effectiveness of a serious game. We tested the methodology by having 320 students play a serious game. The proposed methodology allowed us to infer players' learning outcomes, to assess the game effectiveness levels and to identify issues in the game design.

This research study was partially financed by the Ministry of Education, Culture and Sport of Spain through its 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

Serious games are video games designed for purposes beyond pure entertainment [1]. Serious games are multimedia tools by nature. As a subfamily of videogames, they combine different forms of media (animations, music, text, etc.) to create immersive experiences for players. Their versatility allows them to be used as tools with many applications in different domains. One of the main applications is education, whereby they have become proven learning tools: they are used across several domains with multiple goals and formats, and their acceptance and effectiveness has almost always proven 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 beginning to change. Currently, the widespread use of Virtual Learning Environments (VLE) has allowed for the application of serious games at unprecedented scales. To reach their full potential, serious games should apply 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 has vastly increased the amount of educational data available for analysis. Disciplines such as Learning Analytics (LA) or Educational Data Mining (EDM) study patterns of student interactions to better understand underlying learning processes [5, 6]. This information can be used by different stakeholders for various purposes: from university administrators calculating dropout rates for each class to teachers identifying students at risk of course 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 designates them as rich sources of interaction data. These interactions can be analysed to explore how users play and in the case of serious games to understand how users learn.

The video game industry has been performing these types of analyses on 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 wherein the player feels challenged by the game but is neither bored nor frustrated [9]. GA is used to locate points of gameplay at which players become stuck or quit as well as moments at which a game's mechanics or internal rules fall short. GA is also used to identify clues on ways to fix these problems.

Commercial video games typically non-disruptively collect data from their players through tracking systems that go unnoticed by players [10]. However, according to the literature [11], aspects of serious games are typically assessed through questionnaires completed by players. There is a clear need to combine the emerging disciplines of LA and EDM with the non-disruptive techniques of GA to generate reliable, automated and repeatable assessments of serious games.

Serious game assessments can focus on many outcomes, such as usability, engagement or motivation. However, the learning outcomes are the results most stakeholders wish to obtain from serious games [12]. Learning outcomes have also been the results most frequently assessed when examining recently developed 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 pertains to a lack of methods available to assess serious game effectiveness [14]: teachers, lecturers and policy-makers need to guarantee that serious games are effective enough to be used in the classroom. In this regard, the application of GA techniques to serious games can provide stakeholders with objective and reliable data.

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

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

2. Serious game assessment

Although questionnaires are most commonly used to assess serious games [11], several authors have addressed the implications of using non-disruptive tracking methods for this task. Authors have proposed a set of minimum requirements to enable the automatic assessment of serious games [15] and have addressed the game design implications of combining learning analytics with serious games [16]. The ADAGE project [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” of particular interest in the case of serious games that can be emitted through any video game [18].

Other authors have implemented their own *ad-hoc* analytics, for instance, to analyse players’ steps taken while completing a math puzzle to predict their movements based on current game states [19], to assess learning outcomes by analysing answers to quizzes integrated in a game [20], and to analyse how players progress through learning-language courses to create rich visualizations for teachers [21].

We note that serious game designers must take into account analytics and assessment constraints from a game’s inception and throughout the design phase [15]. Many authors have defined methodologies and guides for designing 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 do not typically address key serious game features, such as ways to deliver knowledge and educational content through gameplay or ways to infer corresponding learning outcomes. Some work has started 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 which lacks reference to concrete methodologies for inferring learning outcomes. On the other

hand, some works have proposed ways to analyse serious game learning outcomes either via general frameworks or ad-hoc analysis, but without addressing the implications of such assessments for game design. We propose combining both approaches in defining a methodology that tackles all phases of serious game development: 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 a serious game (Figure 1). The process starts in the design phase, when the learning goals and target population forms the basis for creating a learning and game design. The combination of these designs is used to create the game, which is then validated through a formative evaluation with a sample of the target population. This process is repeated until the game is fully validated. The game can then be used by the target population (deployment). In the following subsections, we describe each step of the process in greater detail.

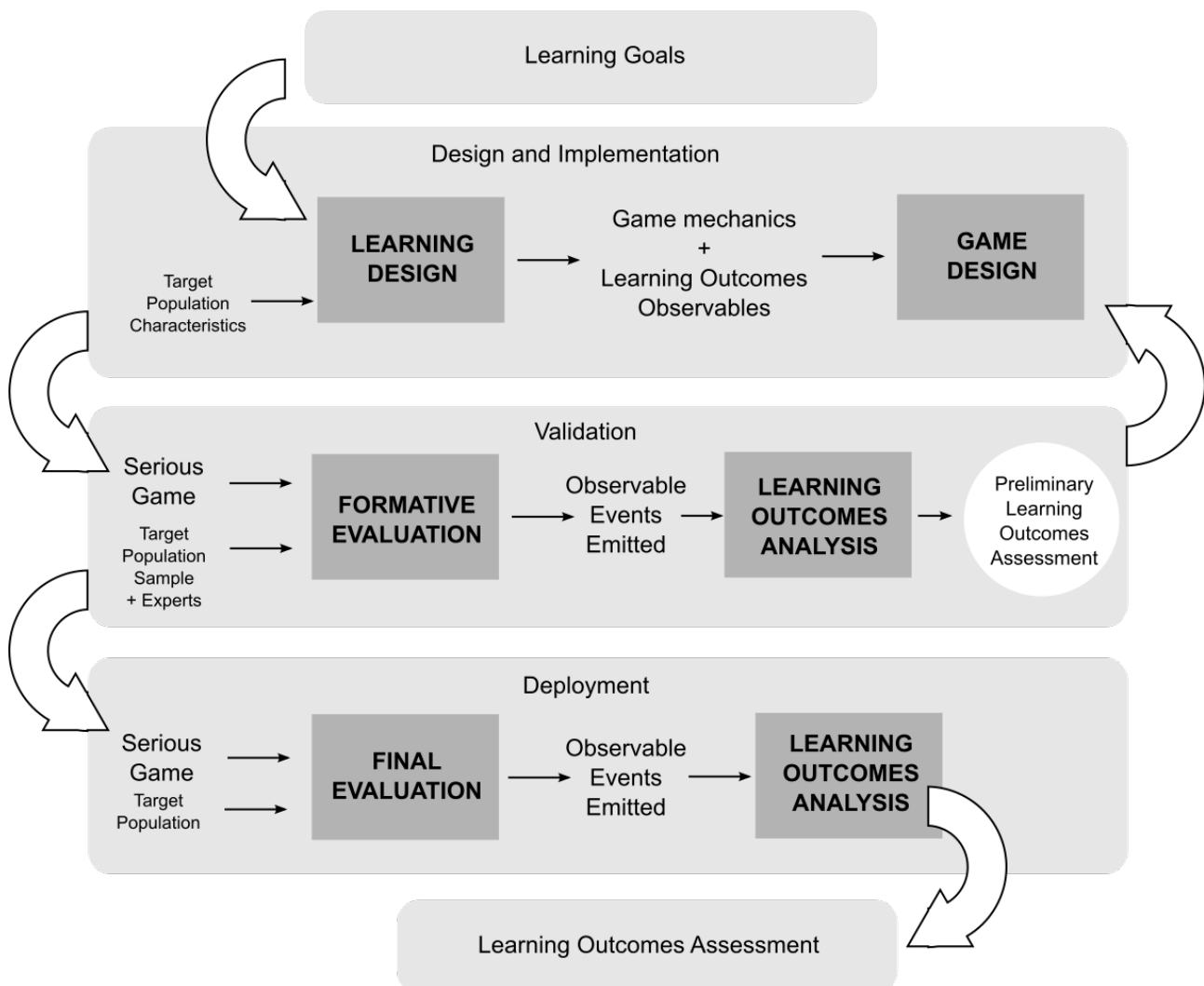


Figure 1. Serious game design and deployment process with learning outcomes assessment.

3.1. Design and implementation

Within the context of our methodology, we define “learning design” as the transformation of learning goals into game mechanics and learning outcomes observables based on characteristics of the target population.

The chosen game mechanics should fulfil two requirements: 1) they should be appropriate for the learning goal content based on models, such as that presented in [27], in which learning mechanics are mapped to game mechanics; 2) gameplay from players should produce learning outcome observables (also termed *events*) that attest to the players’ knowledge or level of skill.

During game design, these constraints, along with many other considerations for a given game (such as art styles, storytelling or technologies) shape the creation of a serious game. Additionally, during this phase, designers must define how a serious game should scaffold the delivery of learning goals. Although there are few concrete methodologies that translate educational theories into game design aspects [28], some authors have proposed models that describe the learning processes of videogames. For instance, in the serious games domain, Kiili proposes the experiential gaming model [29] and problem-based learning [28], both of which are based on an iterative process through which players form a strategy, experiment in the game world, receive feedback, and reflect on the results. In the commercial videogames domain, similar proposals have been made to split experimentation in two sub-steps: experimentation in a “safe game environment”, whereby the

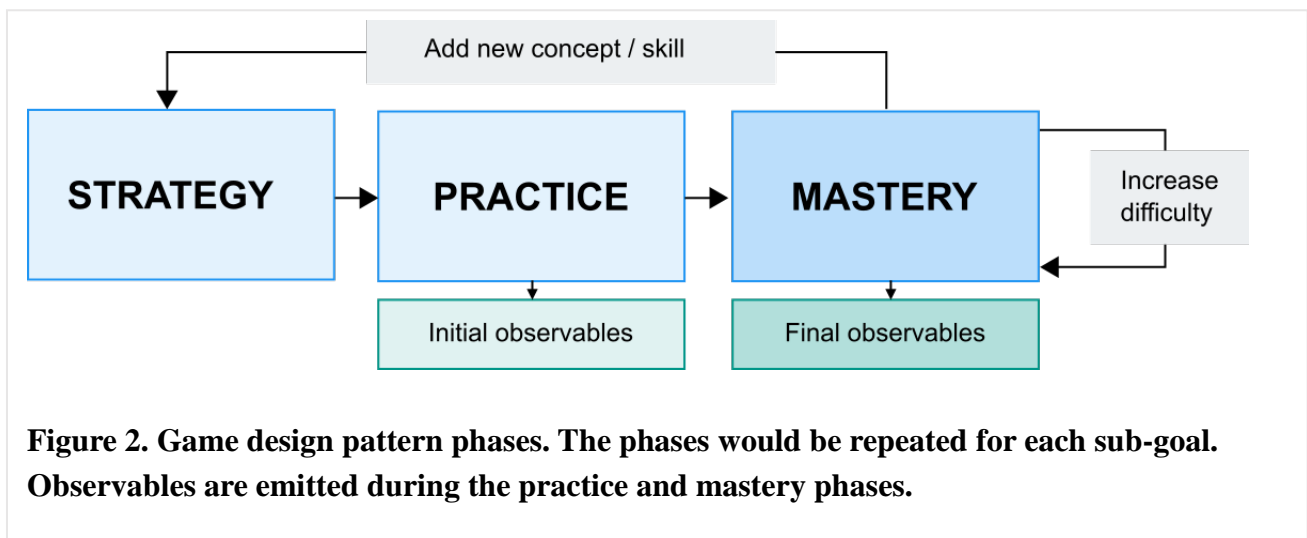


Figure 2. Game design pattern phases. The phases would be repeated for each sub-goal. Observables are emitted during the practice and mastery phases.

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 levels, etc.) [30].

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

- 1) **Strategy:** Players are first introduced to the learning goal. This can involve knowledge they might need in subsequent steps as well as concrete instructions on how to interact with the game world (e.g., through non-interactive scenes or game tutorials). The player receives information on the challenge behind the learning goal and can start identifying ways to tackle the challenge. The strategy phase is similar to initial exposition plus exploration

behaviours that are very common of games.

- 2) **Practice:** players start to apply the knowledge presented in the previous phase. This practice must occur in a game environment in which 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 hypotheses on both the knowledge presented in the previous phase and on the game’s 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.

During this phase, players apply 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 while facing challenges similar to those presented in the practice phase, but with increasing difficulty and higher in-game consequences (“unsafe game environment”).

During this phase, players prove the degree to which they have acquired the targeted skill or knowledge. We can in turn collect final observables that allow us to measure their final progression towards a learning goal.

These three phases can be iterated to deliver multiple learning goals or to deliver a single goal with increasing difficulty while adding a new related concept or skill during each cycle. Additionally, this game pattern optimizes the period in which players occupy the flow zone [9], as it alternates phases in which players learn new things in a safe environment (practice) with phases in which they are challenged to prove their skills (mastery) and through an incremental approach to prevent frustration.

3.2. Collecting observables

Players perform different interactions to advance in a game: they make choices, resolve puzzles, beat bosses, and so on. These events are the core observables on which we perform our learning outcomes analysis. The following principles (many of them shared with general GA) facilitate this analysis:

1. Observables should result in time-stamped events that describe simple interactions between the player and the game [18]. These events should be sent to a central server where all player interactions are stored for later access and analysis.
2. Events sent to the server should be raw interactions rather than opaque scores [18, 31]. For instance, if the mastery phase involves two puzzles, the events to transmit would be the interactions performed to resolve the two puzzles rather than a combined score of the final result. This ensures flexibility, as scores can be later recalculated from interaction data if the subsequent analysis reveals a need to do so.
3. Data collection should be as non-disruptive as possible during gameplay. 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 are stored in a central location, analysis can begin.

3.3. Learning outcome analysis

We store all gameplay interaction events on 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 (as this phase should only contextualize the learning goal). By analysing interactions from the other two phases, we can calculate two assessment scores:

1. **Initial assessment** (*IA*) using initial observables of the practice phase. *IA* estimates the learner's initial level of knowledge. A high value denotes that the player likely possessed the targeted knowledge before starting to play while a low value denotes the opposite.
2. **Final assessment** (*FA*) using final observables of the mastery phase. *FA* estimates the learning outcome. A high value denotes that the player achieved the learning goal while a low value denotes his or her failure to do so.

The specific steps used to transform events into *IA* and *FA* will be different for each serious game. However, they can generally be expressed through formulas that combine data from each interaction. In section 4, we provide details on this process through 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, when *FA*'s value ranges from 0 to 1, a possible value for *FT* could be 0.5, and so we assume that a player achieving an *FA* value of equal to or greater than 0.5 has successfully completed the mastery phase.

For serious games that include multiple learning goals, we can calculate their global *IA* and *FA* values using a weighted average while combining the results of each learning goal: for 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 calculate the global assessment value (A) of 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 the initial and final thresholds as:

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

With these values, we can now estimate learning outcomes and assess serious game effectiveness levels.

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 measure two concrete learning outcomes:

- ***FA* as the player's final score:** We can use *FA* as a score or mark for players when they are considered as students (essentially scoring what they know after playing the game). We must

avoid using *IA* to calculate this marker. Although it represents a player's level of knowledge, using it to calculate final marks would be unfair, as *IA* takes into account mistakes made during the practice phase while a fair grade should only consider what students know at the end of the game and not what they initially ignored.

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

3.3.2. Assessing serious game effectiveness

Within the context of our methodology, we assume that a serious game is effective when we find a positive change in the player's knowledge level. We can determine this change from *IA* and *FA* with respect to *IT* and *FT*. From these values, we can classify each player into a different learning category:

- When $FA \geq FT$, the players have successfully completed the mastery phase and have acquired the targeted skill. Depending on the *IA* value, we can classify players as either:
 - **Learners**, when $IA < IT$: players committed errors during the practice phase, indicating that they did not possess the targeted skill or knowledge before playing the game. However, they ended up being successful in the master phase, suggesting an educational gain during gameplay.
 - **Masters**, when $IA \geq IT$: the players did not commit errors during the practice phase, indicating that they likely already possessed the skill or knowledge before playing the game.
- When $FA < FT$, the players failed the mastery phase and do not possess the targeted skill. Depending on the *IA* value, we can classify players into two different categories:
 - **Non-Learners**, when $IA < IT$: the players also failed the practice phase, indicating that they struggled throughout the game with potentially little or no benefit.
 - **Outliers**, when $IA \geq IT$: 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 by then comparing the total number of players in each category.

When the majority of players are learners, the game is considered highly effective: most players learned something while playing. When the majority are masters, the game is considered to have no learning effect, as most of the players had already possessed the targeted knowledge before playing the game. When the majority are non-learners, the game is considered not effective at all, as most of the players were unable to achieve success at any phase. Finally, a majority of outliers denotes that a game and/or the chosen *FA* and *IA* formulas likely present design flaws.

It is important to note that most serious games will output different results for different populations. A serious game could be highly effective for children of 10 to 12 years of age and not effective at all for children aged 7 to 9 years. The key is to have a well-defined target population during the design of a serious game and to follow a validation process to ensure that effectiveness goals are met.

3.4. Validation and deployment

After applying a serious game along with infrastructure to track its observables in relation to a learning outcomes analysis, we must validate it.

During the validation phase, domain experts and ideally a sample of the target population play a serious game and engage in gameplay that is later assessed through a learning outcomes analysis, yielding preliminary results through a process typically referred to as formative evaluation [32]. This process is iterative and designed to detect ways to fix, polish, tweak or improve a serious game that can range from changing the game mechanics of a learning goal (e.g., when preliminary results suggest low performance) to altering how *FA* and *IA* are calculated (e.g., when experimental results contradict certain game design hypotheses).

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

4. Case study

In the above sections, we presented our methodology for modelling and inferring learning outcomes and effectiveness in serious games. This section describes a case study that illustrates how this methodology works when applied. The case study is based on 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 in regards to learning outcomes and effectiveness levels can be obtained from a serious game developed and analysed through this methodology?

To answer these questions, we used the proposed methodology to implement and analyse “The Foolish Lady”, a serious game¹ based on the homonymous theatre play by Spanish playwright Lope de Vega. In this game, players are presented with several language-related and literature challenges. Its main learning goal is to teach high school students 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 on 320 high school students who 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 by making decisions that affect the overall storyline and final scene. Along the way, they are presented with puzzles and mini-games in which

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

they must apply their knowledge on language and literature. The game is designed to be completed within 30 to 40 minutes.

One of its main learning goals is to teach poetry structure and rhymes, in particular, “redondilla”, a Spanish poetic composition form that uses a specific rhyming scheme and verse length. During the learning design phase, we chose to use point-and-click mini-games as our game mechanic, which relies on drag-and-drop puzzles and option selection in conversations with non-playable, in-game characters. This approach is typical of adventure games, a genre with a track record of proven educational benefits [35]. During the game’s design, we subdivided our goal into the three phases according to our design pattern. Figures 3, 4, and 5 show in-game screen captures representing each of the three phases.

Players are first presented with a textual description of rhymes and of 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 are presented with a mini-game in which they must 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 on the right side of the screen. Once filled, they can check the correctness of the poem they have created by clicking a check button. They can try to do this as many times as they wish until they find the right combination of words: as the *practice phase* of the goal, the results of this mini-game are irrelevant to the final score.

Finally, players are presented with two mini-games (Figure 5). In the first one, players must fight a knight by exchanging rhyming verses. A player can win this battle if he or she selects three correct rhyming replies in a row and loses it if he or she fails three times in a row. The player’s 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 format. Players can answer these questions only once, and both the score and the game protagonist’s marital prospects decrease when they fail. Both mini-games belong to the *mastery phase* of the goal, and therefore the results of these games affect the final score.

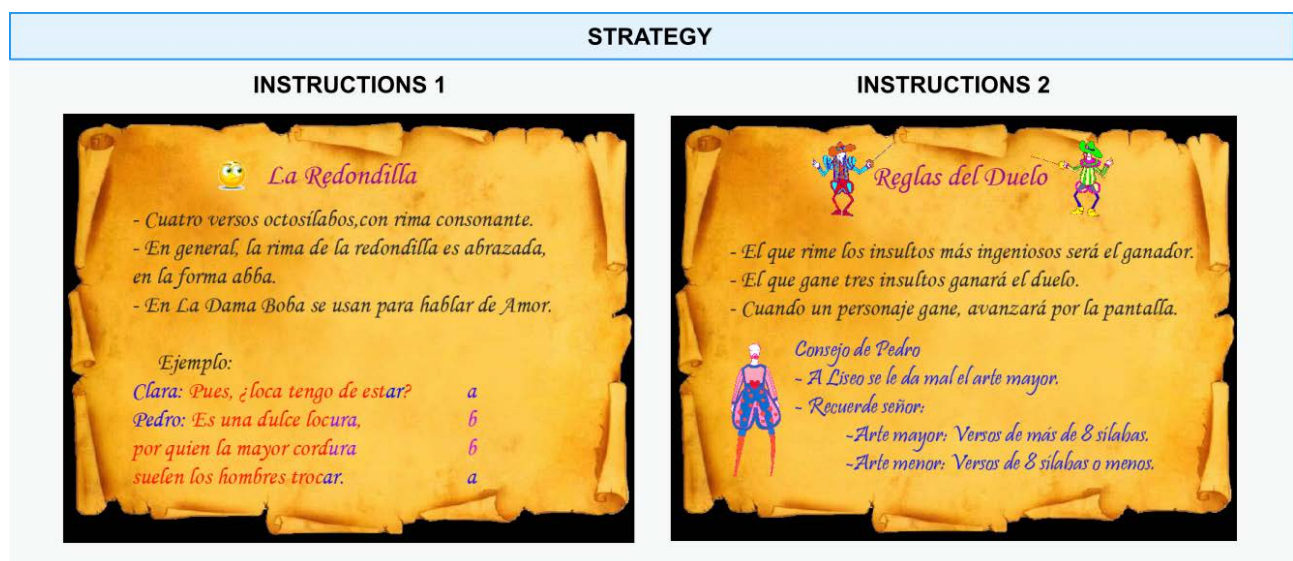


Figure 3. The game presents basic features of the “redondilla” on two screens with textual explanations.

PRACTICE

REDONDILLA PUZZLE

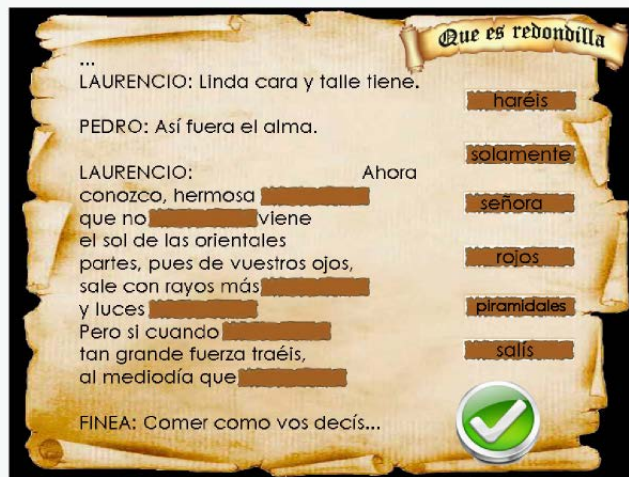


Figure 4. In the first puzzle, players must apply their knowledge of the “redondilla” format. They can try to do this as many times as they wish.

MASTERY

FIGHT MINI-GAME



TEST MINI-GAME



Figure 5. In the last mini-games, players must prove their knowledge. In both cases, the player’s score decreases with each error.

4.2. Collecting observables

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

- Events resulting from a new attempt to complete the “redondilla puzzle”. Every time the player clicks the “check” button and the result is incorrect, a new attempt is made.
- Events resulting from a new attempt to beat the “fight mini-game”. Every time the player loses the fight and restarts the mini-game, a new attempt is made.
- 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 and test mini-games in the mastery phase. For each mini-game, we calculate a score of between 0 and 1:

- **Redondilla Game score (RG):** when A is the observable representing the number of attempts made to solve the “redondilla” puzzle mini-game, RG is computed using $RG = 1 - (\text{MIN}(A - 1, A_{MAX}) / A_{MAX})$, where A_{MAX} is the reasonable number of attempts needed to solve the game. The initial assessment result takes a value of 1 when the player beats the puzzle on the first attempt, i.e., $A = 1$. The initial assessment result takes a value of 0 if the player does not complete the puzzle on any attempt 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 from $FG = \text{MAX}(0, 1 - (\text{MIN}(E, E_{MAX}) / E_{MAX}))$, where E_{MAX} is the maximum number of reasonable errors needed to beat the game.
- **Test Game score (TG):** In the test mini-game, each question has four potential answers, and only one of them is correct. Each answer is given an associated score. The correct answer always has a score of 0, and the remaining answers have scores that correspond to their distance relation 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 listed as jokes. If I is the observable representing the accumulated score of incorrect answers after finishing the test mini-game, $TG = \text{MAX}(0, 1 - I / 4)$, as 4 questions are posed.

We set $A_{MAX} = 3$ and $E_{MAX} = 6$. These values were agreed upon by game designers and educators in consideration of the educational and game challenges that each mini-game presents to players. However, as we wish 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 used.

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

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

For all mini-games, we set the assessment threshold to 0.5, making the IA and FA thresholds 0.5 as well.

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

<i>Initial Assessment / Redondilla game</i> $A_{MAX}=3$ $IA = RG = 1 - (MIN(A - 1, A_{MAX}) / A_{MAX})$		<i>Final Assessment (FA)</i>				
Attempts (A)	RG=IA	<i>Fight game</i> $E_{MAX}=6$ $FG = 1 - (MIN(E, E_{MAX}) / E_{MAX})$		<i>Test game</i> $TG = MAX(0, 1 - I/4)$		FINAL ASSESSMENT $0.5*FG+0.5*TG$
		Errors (E)	FG	Incorrect score (I)	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 for components of FA.

4.4. Case study

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

4.4.1. Experimental design

Before the sample of 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 with the teachers involved in the experiment. The results of this validation allowed us to address some implementation flaws and to improve the gameplay and overall learning design. For instance, two questions from the final mini-game were changed to improve their alignment with the learning goal.

After the validation was conducted, high school students played “The Foolish Lady” for 30 to 40 minutes on PCs under the supervision of a researcher who did not provide any assistance (only brief direction 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*. Students who did not complete a mini-game scored 0. From these variables, we calculated *IA* and *FA* from the formulas presented above. Using their results, we classified each student to a learning category (learner, master, non-learner or outlier) to draw conclusions on the game’s effectiveness.

To gain insight into our methodology, we sought to determine whether we could answer the following case-study questions (CSQ) concerning “The Foolish Lady” serious game:

- **CS1: Did the students acquire the targeted skill by 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 to targeted 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. Thirty-two of the gameplay sessions were corrupted or not completed due to various technical problems that arose during gameplay (power outages, Internet connection issues and computer malfunctions) and were therefore discarded.

The resulting population ($N = 288$) was 44.4% female and 55.6% male. The participants were between the ages of 12 and 16 (with a mean age of 13.70 ± 1.27) and were students at high schools in the Madrid area. Three of these schools were charter or private schools (accounting for 58% of the students), and 4 were public schools (accounting for 42% of the students). In regards to gender, age and school type characteristics, the participants are a representative sample of the student population of Madrid for this age [36, 37].

We also recorded the participants' game habits to classify each student into a player category by evaluating what types of games they played and how often. According to the instrument developed by [38], 14.9% were non-gamers (they never play any video games), 28.8% were casual gamers (they play video games casually for short periods of times), 31.6% were hard-core gamers (they frequently play games such as FPS or MMORPG) and 24.6% were well-rounded gamers (they play all types of games frequently). A more detailed explanation of each category is presented in [38].

5. Results

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

5.1. Game completion

Figure 6 shows the number of players who 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 in player participation (35) occurs between the “redondilla” puzzle and fight mini-game stages.

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

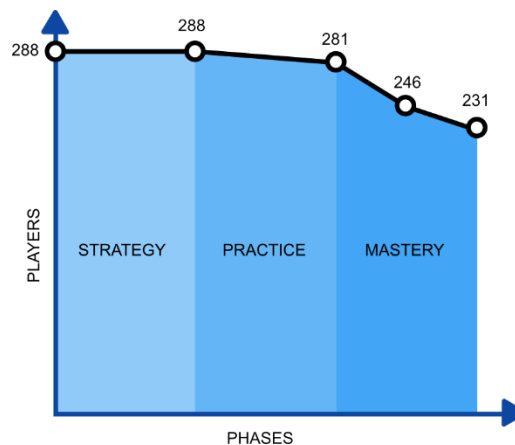


Figure 6: Number of players who accomplished each phase of “The Foolish Lady” game.

5.2. Learning outcomes

To determine whether the students acquired the targeted skill level by 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 and 84.84% of the players who completed the game) scored higher than 0.5 (adequacy threshold set for the game during the design phase) in both FA and IA .

The second part of Case Study Question 1 (CS1) led us to calculate FA and IA across the different

demographic groups: gender (M/F), age (12 to 16) and gaming habits (4 clusters).

To explore whether there were statistically significant differences within each group, we first determined whether the different groups (e.g., males vs. females for division by gender) had a different starting point score. In other words, we needed to determine, for instance, whether males had a statistically significant different *IA* score than females. Regardless of whether such differences existed, we needed to adjust the scores of each group according to their *IA* values before carrying out the analysis.

We therefore performed a one-way analysis of variance (ANOVA) over *IA* to find initial differences across groups. As is shown in Table 2, *IA* showed statistically significant differences for each group, which were especially significant in the case of gender and game habits. Thus, for all of the groups studied, we needed to adjust the initial values through an analysis of covariance (ANCOVA) rather than using the analysis of variance (ANOVA) method, which is generally recommended for similar initial values.

The ANCOVA allowed us to evaluate differences in *FA* scores (dependent variable) across groups (independent variables) by overriding differences in *IA* (that is, by using *IA* as a covariate). Before conducting the ANCOVA analysis, we had to perform standard preliminary checks to confirm that there was no violation of assumptions of normality, linearity, variance homogeneity and regression slope homogeneity [39].

Table 3 shows the ANCOVA results for the 3 independent variables, which present statistically significant differences ($p < 0.05$) among groups by age and game habits but not by gender. This suggests that the dependent variable (*FA*) values differ statistically by player age and gaming habits.

The first ANCOVA [between-subjects factor: age (12 to 16); covariate: *IA* scores] reveals main effects for age $F(4,288) = 7.28$, $p < 0.01$ and a moderate $\eta_p^2 = .094$. According to Table 4, which shows the adjusted means once the effect of *IA* is omitted, the 16-year-old players scored moderately (according to η_p^2) higher ($FA = .766$) than their younger classmates: the 12- and 13-year-olds presented the lowest adjusted means ($FA = .508$).

The second ANCOVA [between-subjects factor: gender (male, female); covariate: *IA* scores] shows no main effects for gender $F(1,288) = .62$, $p = .43$, $\eta_p^2 = .002$. Thus, we can argue that *FA* does not depend on player gender.

A third ANCOVA [between-subjects factor: game-habits (4 clusters); covariate: *IA* scores] reveals main effects for game habits $F(3, 288) = 2.880$, $p = .036$. In this case, the effect of gaming habits ($\eta_p^2 = .030$) was lower than the effect of age. However, the game worked better for well-rounded players and worse for non-gamers.

Independent variable	One-way ANOVAs on <i>IA</i>			
	<i>N</i>	<i>df</i>	<i>F</i>	<i>p</i>
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 the three groups.

Independent variable	ANCOVAs on FA				
	<i>N</i>	<i>df</i>	<i>F</i>	<i>p</i>	<i>Partial η^2</i>
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.

Ind. Variable	<i>Values</i>	ANCOVA		
		<i>N</i>	<i>Adj. Mean*</i>	<i>Std. Err.</i>
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	71	.680	.038
	gamer			
	Hardcore	91	.659	.034

*Adjusted mean using practice phase scores as covariates (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.

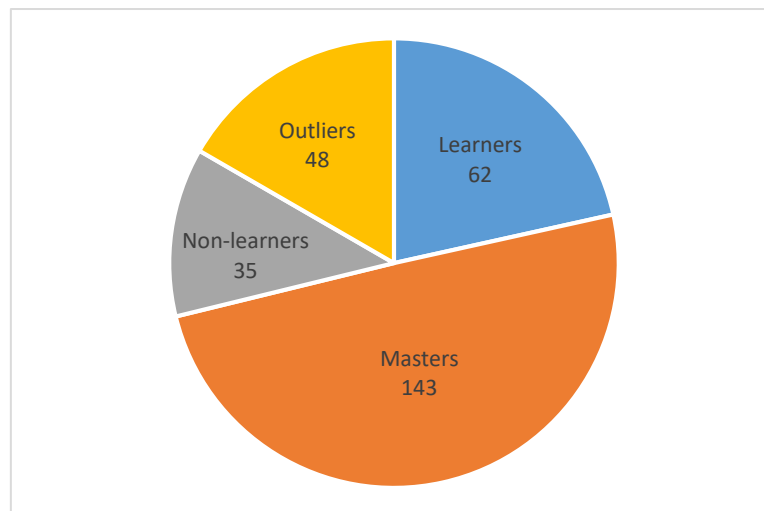


Figure 7. Categorization of players by assessment category.

Figures 8 and 9 show the players grouped by learning category and segmented by age and gaming habits. In all groups, the number of masters exceeds that of the other categories, and especially for the 14-year-old group. For all of the groups, the number of outliers is greater than the number of non-learners, except for the group of students aged 16.

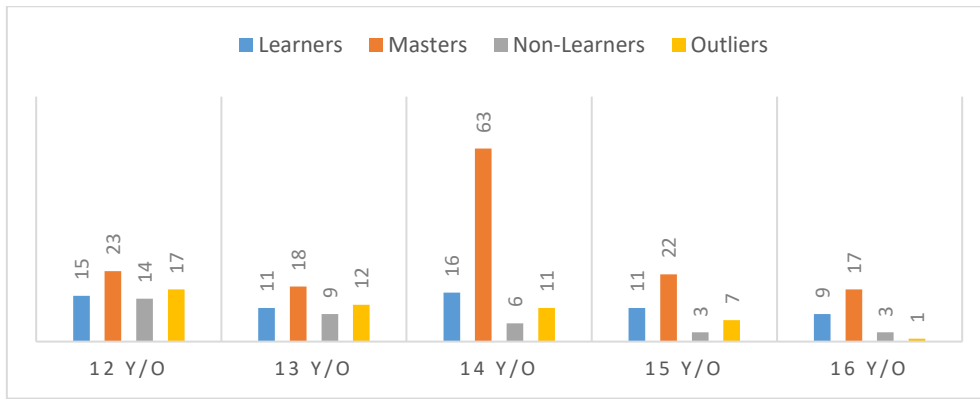


Figure 8. Distribution of players across assessment categories segmented by age.

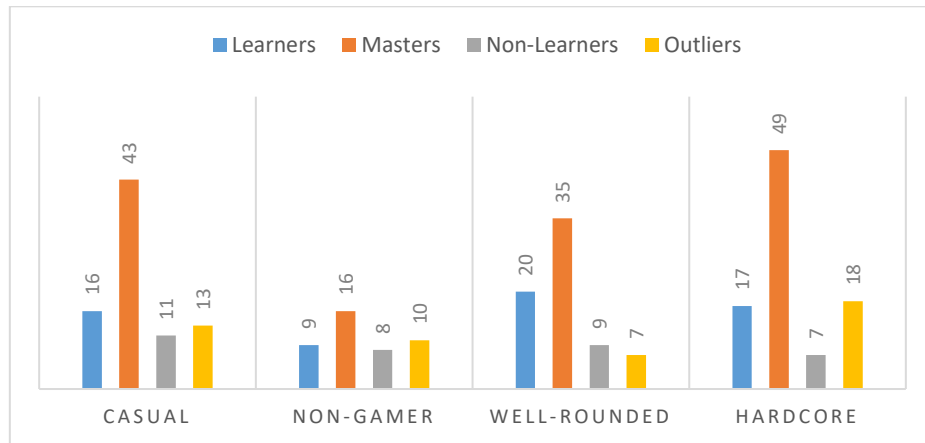


Figure 9. Distribution of players across assessment categories segmented by game-habits.

6. Discussion

In this section, we first present our answers to the case-study questions and then further elaborate on the methodological research questions.

**CS1: Did the students acquire the targeted skill by the end of “The Foolish Lady” game?
Given our demographic variables, were there differences between groups?**

Yes; 80.2% of the students completed the game, which required, by design, a basic understanding of the principles of the learning goal. Regarding the students’ final scores, the ANCOVA analysis (dependent variable: FA; covariate: IA) reveals significant differences when data are segmented by age and gaming habits.

As is shown in a previous section, 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). This seems natural: older students found the game easier.

In terms of gaming habits, the well-rounded gamers generated the best results (Adj. Mean= .680) followed closely by hard-core 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, and so these players’ expertise likely helped them complete the game more effectively. At the other end of the scale, we have non-gamers (Adj. Mean= .554), which supports our hypothesis.

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

No; this is not because the players did not learn, but rather because according to the results, most of them were categorized as “Masters”, i.e., many of them already knew most of the educational content. This could mean that the game was too easy for most of the players. However, we think that an additional problem in the game’s design prevented us from capturing a more accurate IA (and, consequently, a more accurate learning profile): as we wanted to keep the game short—it had to be completed over a standard 40-minute session—the practice phase was made deliberately shorter than the mastery phase. This forced us to limit the practice phase to a single comparatively easy mini-game, which proved insufficient as a full measure of initial knowledge. This flaw went unnoticed during the validation process, as the players initiated tested were domain experts, and it seemed natural to classify them as masters. This oversight shows why serious games must be validated with a sample from the target population and not only by domain experts.

When segmenting groups by age and game habits, there is no particular group for which the game was more effective.

These results do not imply that the game has no value as an educational tool. Students playing this game enjoyed other benefits, such as a measurable increase their motivation to attend 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?

Our use of this methodology forced us to define a clear learning goal from the start and to continue to use it throughout the game development process.

In cooperation with education experts, we clearly defined which role each of the mini-games played in meeting the educational goal. The proposed game-design pattern provided guidance when defining mini-game difficulty, weight and placement measures. The mini-games were implemented in such a way that interactions and events involved in their resolution were clearly identified; and mapping from these events to assessments was also determined early on.

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

We also needed a service to collect all data sent by the tracker. Ours consisted of a REST back-end system that processed HTTP requests describing events, a database for storing these traces, and some Python scripts to query the database. Although we used a customized solution, the serious game could be integrated with any other VLE. This raises new interesting questions regarding the sharing of data between such systems and a serious game, which however fall outside of the scope of this paper.

RQ2. What results in regards to learning outcomes and effectiveness levels can be obtained from a serious game developed and analysed through this methodology?

The identification of relevant educational observables during game design significantly simplified the task of measuring learning outcomes and game effectiveness. These results helped us answer several interesting questions from the case study.

By default, our methodology converts serious games into assessment tools: it relies on clear assessment locations that are associated with both learning design and goals and which are then combined to infer learning outcomes. However, using our design pattern, we can also determine whether students actually learn from playing a game, which is key to assessing a game's effectiveness for a particular population group.

From our case study, we conclude that the initial assessment results are higher than what was expected. The number of outliers generated denotes a design flaw in the practice phase. Such findings, which are based on actual data, are very useful: after altering the game design, the result would be an improved game for the next round of players.

Finally, when student demographic data are available, we can use statistical analyses to identify and characterize those groups enjoying better learning outcomes or game effects. This can help narrow down and characterize the ideal target population of a game. It can also guide changes to adapt the game to other audiences.

7. Conclusions

In this paper, we present a means of structuring the design and assessment of serious games at two levels: inferring learning outcomes and assessing the effectiveness of serious games as educational tools. We think that this will help systematize serious game development and improve several of the methodologies identified in our literature review: our methodology is fully integrated with the production cycle of a serious game (from design to deployment) and involves a non-disruptive assessment as an alternative to questionnaires, the most common assessment method used for serious games. The method poses extra requirements during game development (a tracker in the game engine and a server to collect data), but with the help of today's big data technologies, this is now an affordable task.

We tested our methodology by developing a serious game entitled "The Foolish Lady", which was played by 320 students. The methodology clearly guided the design process and later our analysis of the game's effectiveness. While "The Foolish Lady" proved to be an effective assessment tool (i.e., we were able to assign a mark to each student), it was unable to fully capture the initial knowledge level of the students studied.

One of the conclusions of the experiment is that the design of the practice phase is key to the implementation of an effective serious game. However, balance in the practice phase can be difficult to maintain: the designer wants the player to advance flawlessly while identifying their mistakes to obtain an accurate assessment of their initial knowledge level. In addition, the implementation result derived from the case study (i.e., a tracker and a basic server infrastructure for receiving and analysing traces) is currently being used in the RAGE European Project [40] as an important part of the infrastructure for assessing games.

Although our case study is focused on a serious game designed to deliver knowledge and to teach several skills, we think that 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 control their blood glucose levels could use players' real blood glucose levels (e.g., reading them through a device connected to the game) to determine whether the goal was achieved rather than relying on the use of puzzles or mini-games.

In summary, we conclude that the methodology presented in this paper provides a richer and easily understandable assessment analysis method for serious games. We make one major point that once a game starts sending observable events, everything is automated and all of the assessments are based on how learners interact with the game rather than through the use of traditional out-of-game questionnaires. Additionally, the assessment model is adaptable to researchers' needs, as it is not hardwired to game signals: the way each dependent variable (*FA* and *IA*) is calculated can be changed "a posteriori", allowing constants used in assessment model to be updated when required (for instance, A_{MAX} and E_{MAX} in our case study). Additionally, results obtained via this methodology could complement formal experiments in measuring serious game effectiveness, which remains as an issue to address [41].

We believe that this methodology opens up avenues for future research. In this paper, we limited the students' assessments to 3 particular points in the interest of clarity (the 3 mini-games). In the future, we plan to enrich our game design pattern with more observables for both phases. Additional data will provide us with more information on student progression patterns while affording researchers greater insight into the evolution of the learning process. We plan to go one step further by analysing other gameplay data (such as the time spent on each phase) that may shed light on why some players struggle in certain parts of the game. We also wish to further explore the transformation from game observables into assessment scores by identifying and addressing common patterns of different game mechanics.

Finally, the integration of serious games that follow our proposed methodology within VLEs also raises interesting questions. What standards should be used for such communications? What visualizations should be provided to different stakeholders? Addressing such integration methods will constitute an important step towards realizing the full potential of combining serious games with 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, Thijs 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 “T’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.
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 13 Jan 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 12 Dec 2016
37. Ministerio de Educación (2008) Escolarización y población. <http://www.mecd.gob.es/dctm/ievaluacion/indicadores/2011-e1.2.pdf?documentId=0901e72b810b4d41>. Accessed 12 Dec 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