Applications of *Simva* to simplify serious games validation and deployment

Cristina Alonso-Fernández, Iván Pérez-Colado, Antonio Calvo-Morata,
Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón

***Title*—Applications of *Simva* to simplify serious games validation and deployment**

***Abstract*—Serious games' evaluation and players' assessment is commonly done with experiments where the users play the game and fulfill one or more questionnaires. The tool *Simva* was designed to simplify these complex experiments, which commonly include the collection of game learning analytics data to provide further insight about players' progress and results. We present the latest updates on the tool *Simva* and three applications where it was used to validate serious games using pre-post experiments and collecting game learning analytics data of players' in-game interactions.**

***Index Terms*—Serious Games, Learning Analytics, Evaluation, Assessment**

I. Introduction

This article is an extension of the conference paper presented at LASI-SPAIN 2019 [1] providing further details about the use of *Simva* and fully describing the updated version of the tool, which incorporates new features to simplify its use on more complex experiments.

Serious games (SGs) are applied with multiple purposes, including that of increasing knowledge, raising awareness, or changing attitudes or behaviors. Moreover, due to the interactive nature of SGs, it is possible to track players' actions during gameplay, including not only those related to the achievement of specific design goals, but also any other actions carried out by players that may provide further insight regarding their progress. This latter aspect opens the opportunity to analyze additional aspects of player behavior. In particular, analytics from tracking players has been applied in SGs for two main purposes: first, game design and validation; and second, students' assessment during deployment.

Before moving into their deployment phase, serious games need to be formally evaluated to ensure that their intended purposes are fulfilled. Their formal evaluation will ensure that games are indeed useful for their purposes and that any observed change on players characteristics (e.g. knowledge, attitude) can be attributed to their playing experience [2].

Once games have successfully undergone the evaluation process, it is usually necessary to measure how much of an effect they have had on the players that have used them. In other words, we want to be able to measure how much the serious games have changed whatever they set out to change, such as the knowledge or awareness of its players.

The current approaches used on the literature to evaluate serious games and assess students playing them commonly rely on external paper-based questionnaires and have a correspondingly high cost, as there are no standard solutions that allow both the capture of data from in-games interactions and that gather responses to traditional paper-based evaluation questionnaires. When interactions and questionnaires are gathered separately, it takes additional effort from researchers and educators to link them back together, before attempting to extract useful information from the results.

We present the tool *Simva*, which aims to simplify these costly steps that are often part of serious games' complete evaluation and students' assessment: from interaction data collection and analysis, questionnaires management, and dealing with groups of students, including their anonymization and privacy.

The rest of this paper is structured as follows: Section 2 reviews some of the state of the art regarding evaluation of serious games data collection. Section 3 provides an overview of the tool *Simva* to simplify the validation of serious games, and also describes its latest features. The three following sections exemplify how *Simva* has been used with different serious games to simplify validation and deployment experiments. Section 4 describes the experience of using it with *Conectado*, a serious game to raise awareness about bullying and cyberbullying. In these experiments, *Simva* was used to conduct the experiments to evaluate the game while also collecting interaction data. Section 5 describes the experience with the 15 Objects test, a visual discrimination task. In these experiments, *Simva* was used to evaluate and compare two different versions of the game and two formats (paper-based and computer-based). Section 6 describes the experience with First Aid Game, a game to teach first aid techniques, where *Simva* was used to collect questionnaires and interaction data over both an initial experiment and a later recall experiment. Section 7 discusses the lessons learned from these experiences. Finally, Section 8 summarizes the conclusions of our work.

II. State of the art

The validation of serious games and the measurement of the change they cause on their players has been performed through various methods. Researchers have carried out studies assessing games using methods such as

Manuscrito recibido el día de mes de año; revisado día de mes de año; aceptado día de mes de año.

English version received Month, day-th, year. Revised Month, day-th, year. Accepted Month, day-th, year.

Nombres de los autores, Lugares actuales de trabajo, ciudad, país (email ejemplo@ejemplo.es).

(https://orcid.org/...)

questionnaires, interviews, discussions or logs [3]. These methods can be combined in different ways according to the needs of research and validation. Assessment of the players using these games has also been done with different methods, although there is a trend towards capturing interaction data from gameplays for assessment purposes.

The first step towards evaluating games or assessing the students that play them using their interaction data is to define how interaction data is going to be collected, stored and managed. A common option is to use Learning Record Stores (LRSs), data collection repositories that allow the storage and retrieval of data following the xAPI standard [4]. However, these repositories offer no support to manage important aspects of experiments beyond interaction traces, such as student groupings, privacy, or ancillary data that is part of the experiment but not formatted as xAPI statements.

On the following subsections, we briefly describe the two most common methods to evaluate serious games: pre-post experiments and game learning analytics. Both methods are essential to understand the tool *Simva*, which allows using both together while also managing groups of students.

A. Pre-post experiments

The most common method to evaluate serious games and assess the students who play them is the use of pre-post experiments [3]. These experiments usually comprise three steps:

1. A pre-test: an initial questionnaire which assesses students' characteristic (e.g. knowledge) before the intervention. It can be paper-based or computer-based, and it should be a valid measure of the characteristics that the game aims to change. Therefore, the questionnaire itself must also be formally validated.
2. An intervention: the activity that is intended to change the students' characteristics. In the case of serious games, it will be the gameplay itself, usually from beginning to end. No time should elapse between the pre-test and the intervention, nor between the intervention and the post-test.
3. A post-test: a post-game questionnaire used to assess students' characteristics (e.g. knowledge) after the intervention. This questionnaire has the same requirements as the pre-test, and as such should include the same questionnaire to measure the characteristics, possibly in addition to questions regarding the experience.

The change in the students' characteristics is then measured by comparing the results of the pre-test and the post-test. As the only intervention between the pre-test and the post-test is the gameplay, any significant change (usually an increase in knowledge or awareness) in the characteristic measured by the questionnaires can be attributed to the intervention, which can then be considered as formally validated. Once this process is completed for a serious game, it is ready for deployment in real settings, as it has been proven as useful for its intended goals.

During the deployment phase of games, the educators, managers or researchers applying them will typically want to know how much effect the game is having on their players. For this purpose, it is possible and common to make use of additional pre-post experiments. This, however, may not be the most precise solution, since it is an external measure. In comparison, analytics are internal to the game, and provide time and causality dimensions. However, pre-post experiments can be convenient, and even cost-effective, if the specified educational scenario or research experiment measures a characteristic for which there is no analytics available, such as cases where analytics cannot be provided or there is a lack of resources to include them.

These post-validation questionnaires again provide a measure of the characteristic before the intervention (pre-test) and after the intervention (post-test). For instance, for a game attempting to teach something, the pre-test will show how much players know about the topic before playing, the post-test will report on their knowledge on the topic after playing, and the comparison of both measures will yield how much players have learned with the game. This information will be relevant not only to teachers and educators that wish to measure the effect of games, but also to institutions which can use it to make evidence-based decisions about the application of serious games in their courses.

B. Game Learning Analytics

Besides the external measures provided by the questionnaires, another option to effectively measure the changes on students' characteristics or to obtain insight of students' gameplays on serious games is to analyze their in-game interactions. In the field of Game Analytics for entertainment games, information on player interactions has been collected transparently (in a process called tracking) for many years, primarily with profitability purposes [5]. For serious games, the combination of these Game Analytics techniques with the purposes of Learning Analytics [6] (applied in all kind of learning environments) can be termed Game Learning Analytics (GLA) [7].

GLA data collected from serious games provides information regarding in-game interactions both from an educational perspective and a gaming perspective. This information can therefore be used both to evaluate and improve the game itself, and to gain insights on the progress and results of students, and even to assess them. The information captured from players' interactions can therefore provide a rich insight for a wide set of stakeholders (teachers, managers, educational authorities, researchers, students) and for a variety of purposes (validate game design, students' assessment, improve the game, display real-time feedback, provide overview metrics) [8]–[11].

III. *SIMVA*

The current state-of-the-art regarding the evaluation of serious games and assessment of the students that play them, and our previous experiences in this area, led us to create a tool to simplify these complex and costly processes for any serious game after a short configuration step. For the creation of this tool we first identified several important requirements:

- Data storage and management: collection of questionnaires and game interaction data should be performed in the tool, where results should be available to be downloaded. This avoids dealing with paper-based questionnaires and simplifies linking together all information gathered from each student.

- Student management and anonymization: in-tool creation of groups of students, and assignment of groups of students to questionnaires greatly simplifies keeping track of which questionnaires must be completed by which students. In addition, by issuing students with unique tokens that contain no personal information, and linking all student interaction to those tokens, students can remain fully anonymous without compromising data storage and management.
- Access control and level of completion: access to each questionnaire, and if conducting experiments with serious games, the games themselves should be managed by the tool. This ensures that, for example, students are can only play the game after filling in the pre-test questionnaire. The tool should also provide information regarding player progress and completion of pre- and post-questionnaires, allowing experimenters to gain a quick overview of the progress of the experiment, and to address problems as soon as possible.

By building a tool that fulfills these requirements, we have created *Simva*, which greatly simplifies the validation of serious games, as well as the deployment of games in educational scenarios where additional features such as questionnaires or interaction data collection are desired. *Simva* is capable of managing most of the common requirements for conducting experiments which are guided by questionnaires and where the players' interactions are collected in log files or through a GLA platform. These items include: management of any number of questionnaires, for example the two used for pre-post experiments; of groups of students and users; and of interaction data. Additionally, *Simva* also ensures both privacy and anonymity. With these features, *Simva* can, once suitably configured, be applied to any serious game's evaluation or process of student assessment.

Simva is tightly integrated with LimeSurvey [12], a software that manages questionnaires. *Simva* keeps track of the questionnaires that must be filled by each student, and stores responses for later retrieval. Students are grouped into *classes*, and each class can be assigned a different sequence of activities, such as filling in particular questionnaires or playing specific game versions. For each student in a class, *Simva* generates anonymous 4-letter tokens to be used as their username, instead of any personal identifier that could compromise privacy requirements. *Simva* can connect to a game learning analytics framework, using these tokens to link together all data pertaining to each individual user, regardless of source. Currently supported sources include questionnaires, GLA data, and arbitrary external experimental data uploaded using *Simva*'s web interface.

After generating new tokens for students, *Simva* generates a document that teachers can download and print, to then cut out tokens and hand them out to each student (see **Fig. 1**, bottom image). Teachers have the choice of using this document to write in the names of students next to their tokens for later reference; this allows them to track each student's learning process. However, since no copy of this correspondence between students and tokens is available in the system itself, this still ensures that privacy requirements and regulations are not infringed (e.g. helping to comply with GDPR). This technique also supports performance of recall experiments and longitudinal studies.

Serious games can then be configured to access the specific questionnaires created in *Simva*. A simple configuration file identifies the questionnaires (commonly pre-test and post-test) that the game should access. This is the only configuration required to link the game with *Simva*, therefore, simplifying the use of the tool with any serious game. These questionnaires are linked in *Simva* to one or more classes that should use them. Each class is defined by the set of tokens of its students, and the specific sequence of questionnaires and games that these students will be expected to fill or play. Data collected from each student is stored, and remains linked

Clase ElCaton 1A

☐	Code	Conectado (Pre/Post/Other)			
☐	GYRJ	FINISHED	FINISHED	FINISHED	TRACES
☐	WEFF	FINISHED	FINISHED	STARTED	TRACES
☐	YEYT	FINISHED	NOT FOUND	NOT FOUND	TRACES
☐	ZMBL	FINISHED	STARTED	NOT FOUND	TRACES

Clase ElCaton 1A:					
No.	Nombre	Código			
1		GYRJ	GYRJ	GYRJ	GYRJ
2		WEFF	WEFF	WEFF	WEFF
3		YEYT	YEYT	YEYT	YEYT
4		ZMBL	ZMBL	ZMBL	ZMBL
5		WSFJ	WSFJ	WSFJ	WSFJ

Fig. 1. The *Simva* tool provides access to students' questionnaires and collected interaction traces (top figure) and a ready-to-print list of anonymous tokens that can be handed out to students (bottom figure). Figures retrieved from [22].

together by their anonymous token, including: pre-test, post-test and game interaction data (see **Fig. 1**, top image). All the information is then available from *Simva* to download by authorized users, typically teachers and/or researchers.

An overview of the *Simva* architecture can be seen in **Fig. 2**. *Simva* has been designed to be modular and extensible, and currently has two main modules: surveys (using LimeSurvey) and integration with an analytics framework (the H2020 Rage Analytics platform [13]). *Simva* allows authorized users (teachers or experimenters) to create and manage classes, surveys, games, and link them together to set up experiments.

Classes are groups of students designed to unify student management. When adding a class to *Simva*, teachers or managers set the number of students, and a corresponding amount of four-letter anonymous identifiers (4-letter random tokens such as "FJCD" or "PWNB") are created to hand them out to students, granting them access both to surveys and to the game that is going to be used. When a class is created, *Simva* handles the setup of the corresponding users in both the survey system and the analytics framework, allowing students to send authenticated data with their *Simva*-generated tokens.

A typical experiment has 2 or 3 surveys: a pre-gameplay survey, post-gameplay survey and, if needed, an auxiliary survey. When surveys are created, *Simva* creates the necessary LimeSurvey schema files and uploads them using the corresponding API. Assigning a survey to a class adds the anonymous token identifiers from the class as participants of as valid respondents. Also using the LimeSurvey API, *Simva* can later retrieve survey completion status and use it to allow or deny access to the game itself, preventing the students from playing without answering the survey with a simple status request. When the game ends, a *Simva* API endpoint is available for game-results upload, where logs, scores, or statistics can be appended for later analysis. This can be used to add arbitrary experimental data or metadata for later analysis.

A. *Latest Simva version*

The *Simva* tool has recently undergone an update process to add features and re-structure its internal model of experiments and surveys. The new implementation provides features for both teachers, researchers and students to simplify both the validation and deployment stages of using games in classrooms.

After these updates, users of *Simva* with a role of either teacher or researcher can manage four different entities: groups, studies, tests and activities. Groups of students are managed by the teacher/researcher, who can create them, generate printable anonymous tokens for each group, and assign them to studies. The hierarchy of studies-tests-activities is the main new feature of the new implementation of *Simva*. A study is defined as a full research work to be carried out, and it can involve different stakeholders. For example, a study to validate a serious game may be composed of a validation of the game with students, and another validation with teachers. Studies have one or more associated owners which manage the study, and contain a list of tests. A test can be used to carry out a specific part of the study or a variation of it, for instance, a test may be "Serious game validation with teachers" or "Serious game validation with students", or even "Variant A" vs. "Variant B" for subgroup testing. Tests, in turn, contain a list of activities. Each activity corresponds to a specific task carried out during the study: for instance, an activity may be a questionnaire (e.g. pre-test), a gameplay, etc. **Fig. 3** provides an example hierarchy of a study to validate a serious game with two different tests, each of them containing three and two activities, respectively.

Teachers, like researchers, can manage the creation and assignment of studies, tests and activities. As owners of the studies, they have access to the tests that belong to their studies and, in turn, the activities associated to those tests.

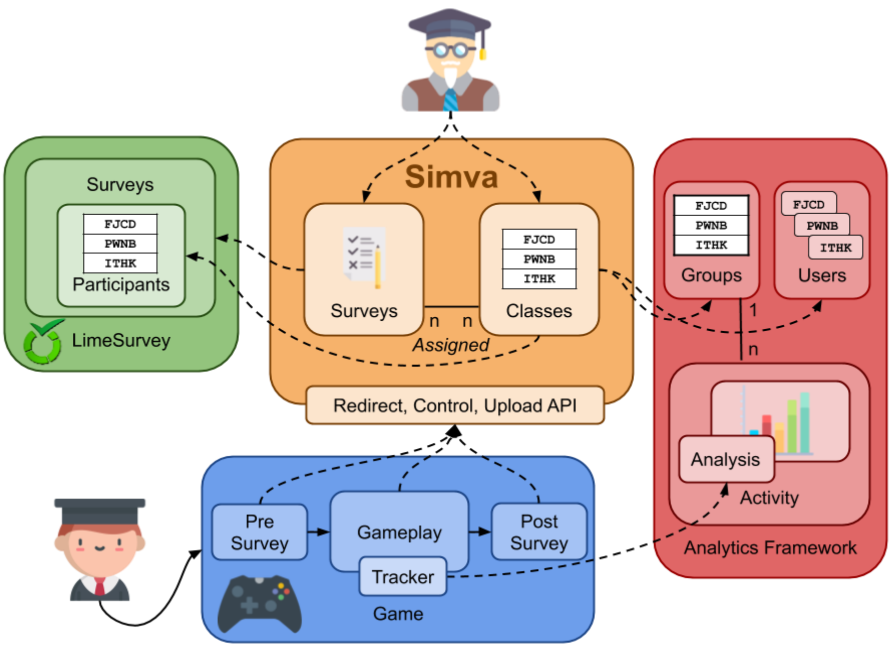

Fig. 2. Overview of the *Simva* architecture managing surveys, classes, and the Analytics Framework. Teachers can access surveys and classes in *Simva*; players typically complete a pre-test, a gameplay and a post-test. All data is sent to *Simva*, linked together using pseudonymous user tokens, and available for later retrieval.

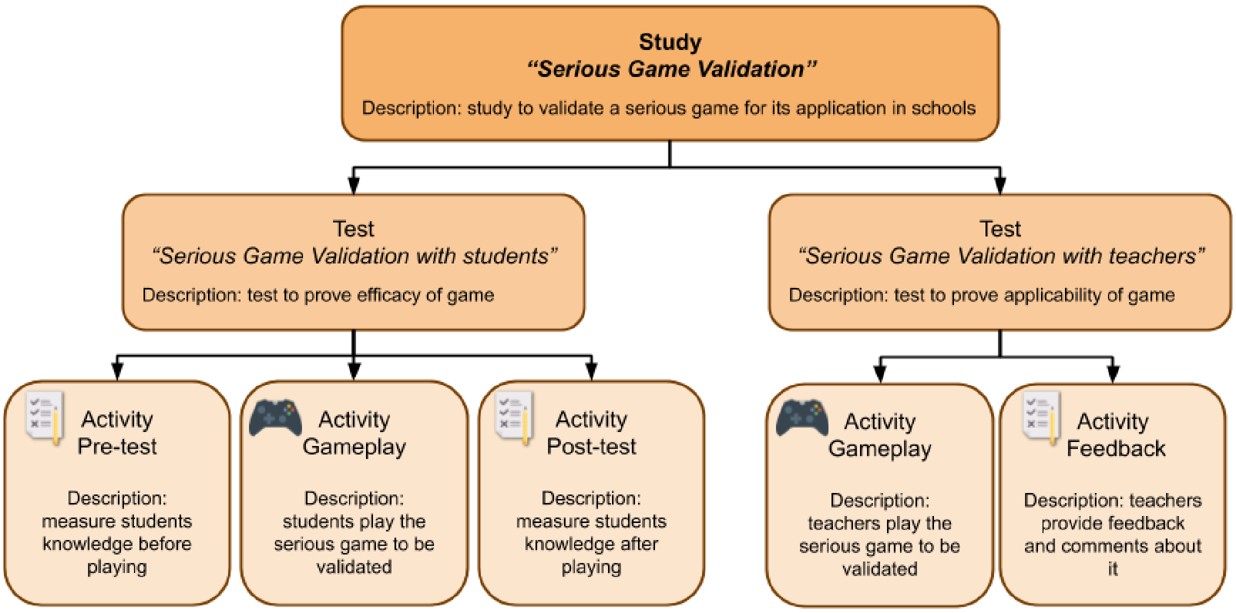

Fig. 3. Example hierarchy of a study for "Serious Game Validation", containing two tests that, in turn, contain three and two activities, respectively. A study with a hierarchy such as this is now possible to be carried out in *Simva*.

This way, teachers can deploy games with associated features such as questionnaires to measure students/players' learning or to provide additional activities together with the gameplay.

Students, on the other hand, simply request information via *Simva* on the studies that they have been added to (as list of activities), and can also request information on each specific activity that they have been assigned to, to know whether it is open or completed, to retrieve a list of links to currently-open activities, or simply to jump to the next activity in a given study. Student activity schedules can be obtained and updated through an API, allowing the user experience to be automated and simplified. For example, the game can connect to this API to make sure that students fill a survey before playing games; or that students complete simple tests between chapters; or even to jump through a set of scheduled mini-games.

Simva has been tested in three case studies, using different serious games, described in the following sections. In the first case study, described in Section 4, *Simva* was used to manage questionnaires and students, and to collect game interaction data, to evaluate a serious game that raises awareness about bullying and cyberbullying. Section 5 describes the experience of using *Simva* to compare two different game versions of a test for active aging. *Simva* managed all questionnaires information and its metadata feature was used to link additional information for each participant. The third experience, described in Section 6, deals with conducting a recall experiment with a game that teaches first aid techniques, and which required collection both of questionnaires and interaction data and using the anonymous identifiers to link both experiments together. *Simva* itself is open-source and available for download and use on GitHub (https://github.com/e-ucm/simva).

IV. Evaluating a Serious Game to Raise Awareness: *Conectado*

Conectado is a graphic adventure video game that aims to raise awareness about bullying and cyberbullying. The game is designed as a tool that teachers can use to kick-start discussions or debriefing sessions on the topics covered in the game with their students. The game places players in the shoes of a transfer student that suffers (in-game) bullying in a new school, and creates a strong shared experience that has proven successful both to increase awareness and to spark debate. So far, the game has been validated through several experiments in high schools with more than 1000 students between 12 and 17 years old, as well as with more than 200 teachers and educational science students [14], [15].

The game validation consisted of pre-post experiments using a formal questionnaire which assesses the players' awareness about bullying and cyberbullying. This questionnaire was used as pre-test and post-test in the experiments to compare players' awareness before and after playing *Conectado*. Additionally, the relevant interactions of the players with the game were collected to further analyze the players' progress, their interactions with other game characters, and their in-game choices and attitudes. This information can provide further insight on how students have used the game and their behavior on bullying and cyberbullying situations such as the ones depicted on the game.

Before conducting the experiments, the main researcher who managed the experience prepared the pre-post questionnaires in *Simva*. This included the following steps: the surveys were registered in *Simva*, and the groups which would be using them were created, with 30 students per group, reserving a few empty student slots as a safety margin. All student slots were identified by their tokens, unique sets of 4 random letters. The list of tokens provided by *Simva* was printed in advance and carried to participating schools. The interaction data was captured by the game and sent to *Simva*, and to the Analytics Server for processing. Therefore, the users created with *Simva* were linked to the ones created in the Analytics System that collected and analyzed user interaction data during the experiments

During the different sessions of the experiments, the main researcher only had to distribute the different printed tokens, one per student. Students then used their tokens to access the game, which had a welcome screen that requested the token to use as identifier for the data which would be sent to the analytics system. At this stage, the game checks in the configuration file for the questionnaires to be used. Then, the game accesses *Simva*, and checks that the assigned pre-test questionnaire is available for the token supplied by the student. If it is available, and not yet submitted, it automatically opens a browser with the initial survey that players must fill in. *Simva* checks that the surveys are correctly configured for the user, as identified by the unique token used to enter the game. If *Simva* indicates that the survey does not exist, or that it is not available for the indicated user, the game will not continue. When the pre-test is completed, the results are sent to *Simva* and users can access the game. After the gameplay is finished, the interaction data is sent to *Simva*. For the post-test, the same checks and process for the pre-test are repeated. If everything is correctly configured, the post-test survey is opened and, when completed, the results are sent to *Simva*.

When the experiments were completed, the main researcher could download the answers to both questionnaires as well as the interaction data from the corresponding *Simva* screen. The different data sources captured from each user (pre-test, post-test and game interactions) were linked together by the unique identifier of each player, facilitating the next analysis step.

With the information gathered in these experiments using *Simva*, the evaluation of the game *Conectado* could be performed, concluding that the game indeed increases the awareness about bullying and cyberbullying, as measured by the pre-post questionnaires. Additionally, analysis of the interaction data captured allowed researchers to extract further information such as times taken to complete the game, progress-over-time, or the different in-game choices and interactions with game characters.

V. Comparing two versions of a serious game for active aging: 15-Objects test

The 15-Objects Test (15-OT) is a visual task that presents 15 overlapping objects that users need to identify as fast as possible. The aim of this test of visual discrimination is to evaluate the slowing of cognitive processing in players that may have Parkinson's disease [16]. The test is carried out with two figures of superimposed images, each with 15 objects, traditionally provided to participants on paper.

For these experiments, in addition to the traditional paper-based version of the test, we developed a new computer-based version of the 15-Objects Test with the same structure and characteristics. This new version was tested with 18 adults [17]. For this test, two different configurations of the 15-OT were created (A and B), each with a different placing of the 15 superimposed objects. To further compare the paper and computerized scores of each participant, as well as the two versions of the game (A and B), participants were randomly assigned to four experimental conditions, balanced by age and gender (see **Fig. 4**). These experiments were a proof-of-concept to test whether the computerized version of this traditional test could be considered equivalent to the paper-based versions when used to research active aging.

For this experiment, the required groups of participants were first created in *Simva*. Then, both questionnaires (pre-test and post-test) were created and managed using *Simva*. These questionnaires were then linked to the groups of participants that were going to use them. All participants needed to complete both questionnaires in different moments according to the conditions shown on **Fig. 4**. Paper-based versions of the test were prepared in advance and handled to participants either at the beginning of the experiment (for participants assigned to experimental conditions II and IV) or at the end (participants assigned to conditions I and III). All participants were provided with their anonymous 4-letter identifiers at the beginning of the experiment. They were asked to write down their unique identifiers on both the pre-post questionnaires, and in the paper-based 15-OT tests as well as to introduce them on their computer-based 15-OT tests.

The experiment was carried out with two different game versions (A and B) depending on the objects used. As seen on **Fig. 4**, some participants (those assigned to conditions II and III) completed version A of the 15-OT test on paper, and version B on a computer. The rest of participants (assigned to conditions I and IV) completed version B on paper and version A on a computer. For the analysis, it was important to know which version of the game each participant was performing on the computer (and therefore which one they were completing on paper). We uploaded this information to *Simva* as additional experimental metadata, linked to each to participants' answers and questionnaires, as this was an early version that did not yet support studies composed of multiple tests. After the additional data was uploaded, the class view clearly showed which version of the game each participant was completing on the computer (A or B). This information was displayed in *Simva* as shown in **Fig. 5**.

After the experiment, participants' responses to both questionnaires were stored in *Simva*, linked together with the information of the game version used in each condition by the unique identifier token provided for students. Researchers could then analyze the questionnaire responses together with the game version to compare both the paper-based and the computer-based versions of the test, as well as to study the

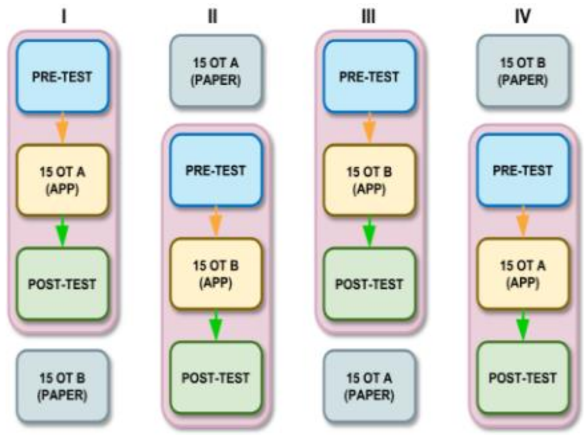

Fig. 4. Four experimental conditions in the *15-Object test* experiments. Participants were randomly assigned to one of these four conditions. Picture adapted from [17].

Fig. 5. A group of participants in the *15-Object test* experiment, showing the game version each one used (A/B). This example shows a possible use of the metadata feature in *Simva*.

equivalence between the two versions of the game used.

The comparison of both game versions was easily carried out, with the aid of the information gathered in these experiments using and *Simva*, also comparing the paper-based and the computer-based versions of the test. Results yielded no significant differences between both game versions, proving their equivalence. Since no significant differences were found in results between the paper-based and the computer-based version of the test, we concluded that the computerized version of the test is a valid and equivalent alternative to the traditional paper-based version.

VI. Collecting GLA data and conducting a recall experiment: First Aid Game

The First Aid Game is a videogame to teach young players life-saving maneuvers in different emergency situations. The game had already been formally validated using a traditional paper-based pre-post setup in a previous experiment. In that validation, the game was even compared with a control group in a theoretical-practical demonstration of the same topics covered in the game [18]. This was, in itself, a good example of the problems that can arise with these types of experiments: after completing the experiments, researchers had to deal with a large number of questionnaires on paper, which were often hard to read, and which had to be painstakingly transcribed to a computer for their analysis.

A new set of experiments was carried out using *Simva*, where data were collected for more than 300 students from 12 to 17 years old [19], [20]. Students completed the pre-post questionnaires assessing their knowledge about first aid techniques, adapted from the questionnaires used on the original validation experiment [18]. For this experience, the game had been rebuilt using a different game technology while maintaining the same design and game mechanics. Nevertheless, these experiments were used to validate that the updated version of the game was still effective at increasing players' knowledge. Additionally, the tracking of in-game interaction data was incorporated to this new version of the game. The interaction data collected while students played the game included game scores, in-game choices and responses, and their interactions with the different game elements. This information helped to follow students' progress and results while they were playing.

In these experiments, the pre-post questionnaires and the groups of students were handled using *Simva*. At the beginning of each session, teachers provided tokens that they had previously downloaded and printed from *Simva* to their students. During the session, teachers wrote down the name of each student next to their token in the printed copies. In this way, teachers were the only stakeholder who had the correspondence between anonymous 4-letter tokens and the students they corresponded to – in a physical medium, safe from online misuse. Tokens could be reused by the same student in the future. No personal information was ever entered in the game. Teachers were encouraged to keep the token print-outs for possible future activities.

A few weeks after completing the training with the game in the school, researchers returned to perform an additional experiment to measure recall of knowledge learned with the game. For this experiment, teachers provided the same token to each student from the paper files they kept (where they had manually written the name of each student next to their assigned token). *Simva* tokens allowed all information from students to be grouped by student while preserving anonymity (from anyone except their teachers), both for the original experiment and the subsequent recall experiment. This greatly simplified the process of analyzing whether students recalled what they had learned, as all the information from their questionnaires and in-game interactions from both set of experiments could be linked together with the anonymous token. **Fig. 6** depicts the experimental setting of these two consecutive experiments using the First Aid Game.

The combination of both sets of experiments helped to measure not only how much students learned while playing, but also how much they remembered a few weeks after the original validation experiment. From their initial knowledge (measured in the pre-test of the original experiment) to their final knowledge (measured in the post-test of the recall experiment), we could determine how much their knowledge had improved with the experience, and in the time in-between (where they could have had other interventions related to the topics covered in the game). In a more fine-grained analysis, recall from the first experience to the second one could be measured by comparing their final knowledge (post-test) on the original experiment with the knowledge they had a few weeks later, before any other intervention (pre-test in the recall experiment). This shows not only that players learn while playing, but also that they are able to recall the things they have learned with the game.

VII. Discussion

Simva aims to reduce the complexity of experiments used when validating and deploying serious games. The three experiences described in this paper (Sections 4, 5 and 6) exemplify how *Simva* has helped to simplify the validation of different serious games as well as the assessment of the students playing them. The evaluation of games has been performed by combining both traditional pre-post experiments with the information collected from in-game interactions in the form of GLA data, after a simple configuration step to link the serious games with *Simva*. Additionally, some features included in *Simva* have also simplified the execution of experiments with complex requirements, such as comparing two versions of a game or conducting a recall experiment.

In the experiments described in this paper, all the

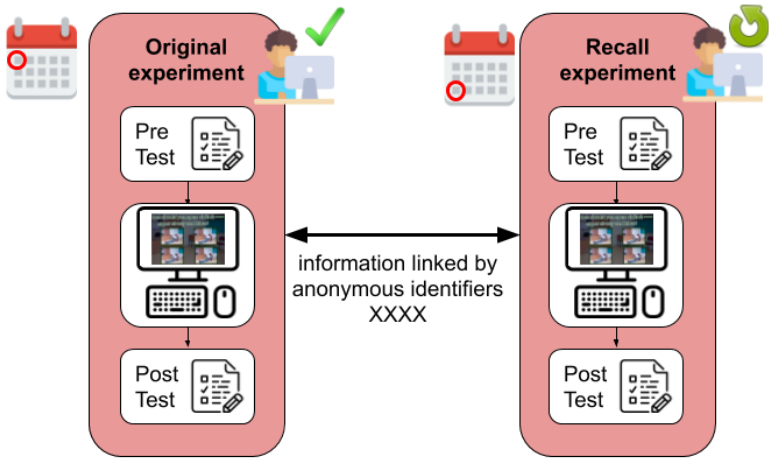

Fig. 6. Experimental setting for *First Aid Game* original and recall experiments. *Simva* simplified linking information between both experiments, and between the different data sources (pre-test, game analytics, post-test) on each experiment.

information gathered from the different sources (questionnaires and game interactions) was collected in *Simva* and linked to each user by their unique anonymous identifier tokens. These identifiers are provided to players by the managers of the activity, who obtained them from the lists of tokens returned by *Simva* when creating the required classes of students. For each participant, researchers have then been able to extract all the information of the experiment: pre-test, post-test, GLA interaction data, and, where applicable, the version of the game played – or the pre-test, post-test, and GLA interaction data from the following recall experiment. All information collected can be used to simplify games validation and deployment by teachers on their own.

From these experiences, we have learnt several lessons that that we consider are key when conducting experiments in real settings to validate serious games, or to deploy them and assess their players:

- **Ensuring user privacy**: to adequately conduct the pre-post experiments, it was essential that the tool we were using to manage students and questionnaires, in this case *Simva*, automatically dealt with and ensured privacy. To effectively ensure privacy, the safest option is to collect and handle no personal information from the experiments. In our experiences, neither *Simva* nor the Analytics System, where interaction data was also being sent to, collected any personal information. Despite ensuring privacy, it still linked together all required information from each student (pre-test, post-test, game interactions, any additional metadata), which was necessary for later analysis. For this purpose, pseudo-anonymization, via the 4-random-letters unique identifier tokens, automatically provided by *Simva* when classes of students are created, has been an effective solution, as privacy is ensured while keeping all student information linked together. For other researchers in similar scenarios, we encourage the use of such a simple anonymization system, which effectively links all the information gathered from each user, simplifying later analysis, while ensuring privacy as the user identifier does not provide any personal information and is the only identification input into the system.
- **Collecting different data sources**: the collection of questionnaires done with *Simva*, as automatically launched from the game and performed online collecting the results without any effort, reduces the time and cost of carrying out pre-post experiments, as well as the use of paper and printer supplies. An additional option was also available during the experiments to collect the information offline. With this option, all data was stored in the computers were students were playing, to be later manually collected by researchers in case of network connection problems. All the interaction data was also stored and linked with the questionnaires online and offline. The option to upload additional metadata to *Simva* as well as the possibility of linking the information from several experiments together has also been very useful, as it simplifies the later analysis of each users' data, regardless of source. We encourage researchers to consider similar options to link together their different data sources as it greatly simplifies the later steps of an experiment's analysis.
- **Moving from pre-post experiments to GLA**: although, in the three cases presented, we have used the traditional pre-post experiments to evaluate serious games' efficacy and assess students playing the game, we consider that the information extracted from in-game interactions is also essential, and research should move towards always including this type of information. In our case, we have used the xAPI-SG Profile [21] as the data collection standard for in-game interactions. This Profile describes a common set of interactions that describe the main events found in serious games, and a specific set of verbs and activity types to use when storing them as xAPI statements. We used different trackers for each game, but since they all sent interaction data using the same xAPI-SG Profile, all of them could be analyzed and displayed using the same tools. We also recommend researchers to use this or other standard when collecting interaction data as it simplifies data reuse and integration in larger systems, including the

collection process itself [11].

We consider that the combination of these lessons learned from our work, collecting different data sources from traditional pre-post experiments to more informative GLA data while ensuring privacy, can benefit the execution of experiments to evaluate serious games and assess the students who play them. The tool *Simva* has been effective in the described experiments as it has greatly simplified the execution of the three case studies. New features added after these experiments can simplify the execution of even more complex, long-term experiments and deployments by using its new hierarchy of studies, tests and activities.

VIII. CONCLUSIONS

Increased adoption of serious games in education builds upon their previous formal validation. Despite their drawbacks, pre-post experiments are still one of the most common evaluation methods for serious games and for the assessment of students who play them. Together with these classical experiments, the use of game interaction tracking has also increased, providing richer information based on the in-game actions of students/players. The combination of pre-post experiments and collection of game learning analytics data make the formal validation of games complex and costly. Simplifying these experiments and making them more user-friendly and reducing their costs both in time and effort can greatly improve the application of games in real settings, simplifying their evaluation and deployment, and increasing their application including assessment of players.

The tool *Simva* was created for this purpose to simplify all the issues that need to be dealt with when conducting experiments to validate serious games. *Simva* manages both questionnaires and groups of players, addresses privacy concerns, allows the collection and linking of information from different data sources (both questionnaires and in-game interactions), and includes additional features for requirements such as adding arbitrary experimental metadata or carrying out recall experiments. Additionally, *Simva* has been recently updated to extend and improve management of complex or multi-part experiments. By adopting a three-layer hierarchy of studies, tests and activities, experiments that include complex setups, span multiple stages, or involve several stakeholders are much easier to carry out. As future work, we plan to validate this latest *Simva* extension on new experiments with these characteristics.

The three specific applications of *Simva* described in this paper show how the tool has been used in real settings and for different goals related to the evaluation of serious games and the assessment of the students playing them. We have also presented several lessons learned from these experiences, which we hope will contribute to further research on this area.

ACKNOWLEDGMENT

This work has been partially funded by the Regional Government of Madrid (eMadrid P2018/TCS4307), by the Ministry of Education (TIN2017-89238-R) and by the European Commission (RAGE H2020-ICT-2014-1-644187, BEACONING H2020-ICT-2015-687676, Erasmus+ IMPRESS 2017-1-NL01-KA203-035259).

REFERENCES

[1] C. Alonso-Fernández, I. J. Perez-Colado, A. Calvo-Morata, M. Freire, I. Martinez-Ortiz, and B. Fernandez-Manjon, “Using Simva to evaluate serious games and collect learning analytics data,” in *LASI-SPAIN*, 2019.

[2] C. S. Loh, Y. Sheng, and D. Ifenthaler, *Serious Games Analytics*. Cham: Springer International Publishing, 2015.

[3] A. Calderón and M. Ruiz, “A systematic literature review on serious games evaluation: An application to software project management,” *Comput. Educ.*, vol. 87, pp. 396–422, Sep. 2015.

[4] ADL, “xAPI Lab.” [Online]. Available: http://adlnet.github.io/xapi-lab/. [Accessed: 03-Jun-2016].

[5] M. El-Nasr, A. Drachen, and A. Canossa, *Game Analytics: Maximizing the Value of Player Data*. London: Springer London, 2013.

[6] P. Long and G. Siemens, “Penetrating the Fog: Analytics in Learning and Education,” *Educ. Rev.*, pp. 31–40, 2011.

[7] M. Freire, Á. Serrano-Laguna, B. M. Iglesias, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón, “Game Learning Analytics: Learning Analytics for Serious Games,” in *Learning, Design, and Technology*, Cham: Springer International Publishing, 2016, pp. 1–29.

[8] V. Shute and M. Ventura, “Stealth Assessment,” in *The SAGE Encyclopedia of Educational Technology*, 2455 Teller Road, Thousand Oaks, California 91320: SAGE Publications, Inc., 2013, p. 91.

[9] A. R. Cano, B. Fernández-Manjón, and Á. J. García-Tejedor, “Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities,” *Br. J. Educ. Technol.*, vol. 49, no. 4, pp. 659–672, Jul. 2018.

[10] M. Muratet, A. Yessad, and T. Carron, “Understanding Learners’ Behaviors in Serious Games,” in *Advances in Web-Based Learning - ICWL 2015*, vol. 9412, F. W. B. Li, R. Klamma, M. Laanpere, J. Zhang, B. F. Manjón, and R. W. H. Lau, Eds. Cham: Springer International Publishing, 2016, pp. 195–205.

[11] C. Alonso-Fernández, A. Calvo-Morata, M. Freire, I. Martinez-Ortiz, and B. Fernández-Manjón, “Applications of data science to game learning analytics data: a systematic literature review,” *Comput. Educ.*, 2019.

[12] The LimeSurvey Project Team, “LimeSurvey,” 2013. [Online]. Available: https://www.limesurvey.org/.

[13] eUCM Research Group, “RAGE Analytics,” 2016. [Online]. Available: https://github.com/e-ucm/rage-analytics. [Accessed: 13-Nov-2016].

[14] A. Calvo-Morata, D. C. Rotaru, C. Alonso-Fernández, M. Freire, I. Martinez-Ortiz, and B. Fernandez-Manjon, “Validation of a Cyberbullying Serious Game Using Game Analytics,” *IEEE Trans. Learn. Technol.*, pp. 1–1, 2018.

[15] A. Calvo-Morata, M. Freire-Moran, I. Martinez-Ortiz, and B. Fernandez-Manjon, “Applicability of a Cyberbullying Videogame as a Teacher Tool: Comparing Teachers and Educational Sciences Students,” *IEEE Access*, vol. 7, pp. 55841–55850, 2019.

[16] B. Pillon *et al.*, “Cognitive slowing in Parkinson’s disease fails to respond to levodopa treatment: The 15-objects test,” *Neurology*, vol. 39, no. 6, pp. 762–762, Jun. 1989.

[17] M. C. D.-M. Dan-Cristian Rotaru, Sara García-Herranz, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón, “Using Game Technology to Automatize Neuropsychological Tests and Research in Active Aging,” *GOODTECHS 2018 - 4th EAI Int. Conf. Smart Objects Technol. Soc. Good*, 2018.

[18] E. J. Marchiori, G. Ferrer, B. Fernandez-Manjon, J. Povar-Marco, J. F. Suberviola, and A. Gimenez-Valverde, “Video-game instruction in basic life support maneuvers,” *Emergencias*, vol. 24, no. 6, pp. 433–437, 2012.

[19] C. Alonso-Fernández, A. R. Cano, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, “Lessons learned applying learning analytics to assess serious games,” *Comput. Human Behav.*, vol. 99, pp. 301–309, Oct. 2019.

[20] C. Alonso-Fernández, R. Caballero Roldán, M. Freire, I. Martinez-Ortiz, and B. Fernández-Manjón, “Predicting students’ knowledge after playing a serious game based on learning analytics data: A case study (in press),” *J. Comput. Assist. Learn.*, 2019.

[21] Á. Serrano-Laguna, I. Martínez-Ortiz, J. Haag, D. Regan, A. Johnson, and B. Fernández-Manjón, “Applying standards to systematize learning analytics in serious games,” *Comput. Stand. Interfaces*, vol. 50, pp. 116–123, 2017.

[22] I. J. Perez-Colado, C. Alonso-Fernández, A. Calvo-Morata, M.

Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, “Simva: Simplifying the scientific validation of serious games,” in *9th IEEE International Conference on Advanced Learning Technologies (ICALT)*, 2019.

Cristina Alonso Fernández obtained her Bachelor in Computer Science and her Bachelor in Mathematics for the Complutense University of Madrid in 2016. A year later, she finished the Master in Data Mining and Business Intelligence, also for the UCM. Since September 2016, she is part of eUCM, where she has worked for the H2020 Beaconing Project. She is currently doing her PhD in Computer Science. Her research interests include educational videogames and application of data analysis and data mining for their improvement.

Iván José Pérez Colado received a Bachelor in Software Engineering for the Complutense University of Madrid in 2014. Two years later, he completed the Master in Computer Science, also at the UCM. As his Master Thesis was uAdventure, a new implementation of eAdventure in Unity 3D**,** Dr. Baltasar Fernández Manjón offered Iván the chance to continue the development of uAdventure being a researcher of the eUCM Group. He also started his PhD in Computer Science that same year. Six months after that, Iván became part of the H2020 Beaconing Project**.** His research interests include **e**ducational videogames and the authoring tools used to create them**.** Due to his work in Beaconing, the study of Learning Analytics techniques and standards is one of his main research areas.

Antonio Calvo-Morata obtained his bachelor in Computer Science for the Complutense University of Madrid in 2014. In 2017, he completed the Master in Computer Science, also in the Complutense University. He is currently doing his PhD in Computer Science. Antonio has been part of the research group e-UCM since 2014, as a contract researcher for projects eMadrid and H2020 RAGE. His research interests include the study of educational videogames and their application in schools, as well as the use of Learning Analytics techniques to improve their efficacy and their validation as an educational tool.

Manuel Freire has a PhD in Computer Science from the Universidad Autónoma de Madrid (UAM). He is interested in Information Visualization, Human-Computer Interaction, Online Learning, Serious Games, and Plagiarism Detection. He performed a 2008 post-graduate Fulbright scholarship in the Human-Computer Interaction Lab (HCIL-UMD), working with Ben Shneiderman and Catherine Plaisant. In 2010, he became a member of the e-UCM group in the Universidad Complutense de Madrid (UCM), where since 2013 he is an Associate Professor.

Iván Martínez-Ortiz works as Associate Professor in the Department of Software Engineering and Artificial Intelligence (DSIA) at the Complutense University of Madrid (UCM). He has been assistant to the Vice-Rector of Technology at UCM and Vice-Dean for Innovation in the Computer Science Studies. He has been Lecturer in the Computer Science School at UCM in the Computer Science School at the Centro de Estudios Superiores Felipe II. He received a Bachelor in Computer Science (first in the Dean’s List “premio extraordinario”) and a Master and PhD in Computer Science from the UCM. His research interests include e-learning technologies and the integration of educational modeling languages, serious games and e-learning standardization.

Baltasar Fernández-Manjón received the PhD degree in physics from the Universidad Complutense de Madrid in 1996. He is a Full Professor of computer science at UCM and director of the e-UCM e-learning research group. He has the Honorary Complutense-Telefonica Chair on Digital Education and Serious Games. His research interest is focused on the applications of ICT in education and in serious games and educational simulations applied to different domains (e.g. medicine, education). He is also working in game learning analytics and the application of e-learning standards to the integration of those technologies in e-learning systems. He is a senior member of the IEEE.