
Routledge
Taylor & Francis Group

Interactive Learning Environments

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/nile20

# Evidence-based evaluation of a serious game to increase bullying awareness

**Cristina Alonso-Fernández , Antonio Calvo-Morata , Manuel Freire , Iván Martínez-Ortiz & Baltasar Fernández-Manjón**

**To cite this article:** Cristina Alonso-Fernández , Antonio Calvo-Morata , Manuel Freire , Iván Martínez-Ortiz & Baltasar Fernández-Manjón (2020): Evidence-based evaluation of a serious game to increase bullying awareness, Interactive Learning Environments, DOI: 10.1080/10494820.2020.1799031

**To link to this article:** https://doi.org/10.1080/10494820.2020.1799031

Published online: 03 Aug 2020.

INTERACTIVE LEARNING ENVIRONMENTS
https://doi.org/10.1080/10494820.2020.1799031



# Evidence-based evaluation of a serious game to increase bullying awareness

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

Department of Software Engineering and Artificial Intelligence, Faculty of Computer Science, Complutense University of Madrid, Madrid, Spain

**ABSTRACT**
Game Learning Analytics can be used to conduct evidence-based evaluations of the effect that serious games produce on their players by combining in-game user interactions and traditional evaluation methods. We illustrate this approach with a case-study where we conduct an evidence-based evaluation of a serious game's effectiveness to increase awareness of bullying. In this paper, we describe: (1) the full process of tracking in-game interactions, analyzing the traces collected using the standard xAPI-SG format, and deriving game learning analytics variables (to be used as *evidences*); and (2) the use of those variables to predict the increase in bullying awareness. We consider that this process can be generalized and replicated to systematize, and therefore simplify, evidence-based evaluations for other serious games based on the interaction data of their players.

**ARTICLE HISTORY**
Received 5 February 2020
Accepted 15 July 2020

**KEYWORDS**
Serious games; learning analytics; stealth assessment; technology-enhanced learning; game-based learning; e-learning

## Introduction

Serious Games (SGs) are defined as games whose purposes go beyond simple entertainment, such as teaching, training, or changing perceptions. For instance, serious games have been used to raise awareness about social problems or to effectively change players' attitudes or behaviors (Peng et al., 2010; Xu et al., 2014).

The traditional methodology to evaluate serious games uses paired external formal questionnaires to assess players: one before the application of the game (pre-test) and one after the gameplay is finished (post-test). Results of both questionnaires are then compared to evaluate the game's effect on its players and, therefore, whether it achieves its intended goals (Calderón & Ruiz, 2015).

Given the availability of in-game interaction data, data-based approaches can gain deeper evaluation insights than those which only rely on paired questionnaires. Game Learning Analytics (GLA) brings together the fields of Learning Analytics (in education) and Game Analytics (in game industry), and is defined as the tracking, collection and analysis of data from the interaction of players with serious games for several purposes, such as improving the game design, understanding players' mental processes, or assessing their learning (Alonso-Fernández, Calvo-Morata, et al., 2019; Freire et al., 2016). On a previous work, we started to investigate how to combine both traditional formal and widely-accepted methods (pre-post experiments) and more recent and powerful techniques (Game Learning Analytics) to assess players' characteristics using a serious game (Alonso-Fernández et al., 2020). This latter approach poses the basis for evidence-based serious games evaluation.

**CONTACT** Cristina Alonso-Fernández calonsofernandez@ucm.es

This paper presents a case-study of an evidence-based evaluation of the effect of *Conectado* using GLA and data mining techniques. *Conectado* is a serious game to raise awareness about bullying and cyberbullying for players between 12 and 17 years old. The game is designed as a tool for teachers to use in their classrooms to spark discussions on this topic after all players/students have shared the common experience of playing, and has been validated through several experiments in schools (Calvo-Morata et al., 2019, 2020). The evidence-based evaluation of the effect of the game is performed in two steps: (1) **obtaining evidences from the serious game:** describes the collection of data from the interaction between players and games, including its representation (in our case, using the standard xAPI-SG Profile), and the processing and analysis of the resulting data to obtain relevant GLA variables; and (2) **using those evidences to predict the effect of the serious game:** the GLA variables derived from xAPI-SG traces are used as input for prediction models, which, once trained, can predict how the characteristics that the game is built to change (in our case, bullying awareness) will effectively change. This evidence-based evaluation aims to simplify players' assessment in further deployment of the games, where the effect on players could be calculated based solely on the actions they take in the game (*evidences*).

The rest of this paper is structured as follows: the "Related work" section presents previous research on assessment of learning for serious games; the "Materials and methods" section describes the game and methodology used in this case-study, including the experiments that we have carried out to collect interaction data from *Conectado*, the variables we derived from those traces, and the models we chose to predict increase of bullying awareness based on in-game actions. Prediction results are then presented in the "Results" section. Finally, the "Discussion" analyzes our process and its results section, while the "Conclusion" section summarizes the contributions of this work.

## Related work

The application of serious games in different domains has increased the interest towards the integration of players' assessment in the videogame itself (Shoukry et al., 2014). The method to conduct the assessment of learning has been traditionally based on external measures (e.g. pre-post tests), but has recently been shifting towards more evidence-based approaches. For instance, *stealth assessment* (Shute et al., 2017; Shute & Ventura, 2013) aims to embed the assessment in the game, collecting evidences in a non-disruptive manner while the game is in play. These evidences are then used to update a game-based model that informs the results of players' assessment.

In a literature review on the uses of learning analytics data for assessment in serious games, Liu *et al* pointed out the need of more studies that combine different data sources, for example traditional measures, such as questionnaires, with more dynamic data, such as in-game metrics. They also noticed the lack of standard procedures to guide researchers on how to use in-game data (Liu et al., 2017). These conclusions have been pointed out by other researchers, highlighting the need of more research on how serious games can be effectively used for assessment, and which characteristics of such games contribute or detract to their validity for assessment (Kato & Klerk, 2017). On a previous literature review on data applications in serious games, we also found out that assessment was the main purpose of such applications, but that more research was required to establish general approaches, and that the sample sizes used in such studies should be increased (Alonso-Fernández, Calvo-Morata, et al., 2019).

## Materials and methods

As previously mentioned, the game to be evaluated is *Conectado*, a serious game to raise awareness about bullying and cyberbullying. In *Conectado*, players play in first person as a student that transfers into a new school, and, during the first week, becomes increasingly bullied by classmates. Those aggressions happen both in the school and at home, where the bullying continues via social media (which makes it cyberbullying). The game has a linear flow and, depending on the actions taken, such as mentioning the problem with the character's parents or teachers, players will reach one of the three different game endings. By design, player's choices only have an immediate effect on the next dialogs but do not affect the main storyline until just before the ending. This ensures that all players will go through all the situations represented in the game, while still experiencing their actions as meaningful, even while they have minimal effect on the overall flow of the game. Linear play also makes all playthroughs of comparable length, and provides all players with a common experience for their in-class post-game discussions.

This case-study with *Conectado* comprises two phases. In the *experiment phase*, we have collected data from pre-post questionnaires, together with *Conectado* in-game interaction data as represented using the xAPI-SG Profile. In the *analysis phase*, we have processed those xAPI-SG traces to derive a set of GLA variables, which we have then used as input for prediction models to predict the increase in bullying awareness. The actual increase in bullying awareness, to compare the predictions of the models with, is obtained by comparing the scores of each players' pre- and post- tests as gathered in the experiment phase.

Figure 1 depicts how this two-step methodology is organized. Pre-post questionnaires are collected first, through experiments with actual students in their classrooms, and the results of these experiments are later used to derive the target variable (increase in bullying awareness) of the prediction models (green lines). In-game interaction traces for *Conectado* are automatically collected using a game tracker, that is, a reusable component that allows developers to communicate the SG with an analytics platform. These traces are processed to obtain the values of the GLA variables (which have been previously defined based on knowledge of the game, or by relying on a default set of variables), and these variables are then used as input for the prediction models (red lines). We have used Python both to process the xAPI-SG traces and to train and use the prediction models.

## *Collecting interaction data from Conectado*

The data used in the case-study was obtained from $N = 1109$ participants (ages 12–17) from 11 schools around Spain. In all experiments, participants completed a pre-test, a gameplay of *Conectado*, and a post-test, in that order. Minimal time elapsed between the gameplay and either of the two tests, and the complete sessions lasted a total of around 50 minutes, fitting in an average-length lecture session in Spain's schools.

The pre-test and the post-test both assess bullying and cyberbullying awareness before and after playing *Conectado*. The set of questions included in both tests derive from multiple formal and widely-accepted questionnaires that have been demonstrated effective in the school population of Spain (Álvarez García et al., 2013; Garaigordobil & Aliri, 2013; Ortega-Ruiz et al., 2016). In total, the pre-test and the post-test included 18 7-point Likert questions, eliciting how much players agree with each of 18 statements on bullying and cyberbullying. The questionnaire has a Cronbach's alpha of 0.95. The score of each test is calculated as the mean of all answers; therefore, possible test scores range from 1 to 7. Questionnaires and interaction data were managed with *Simva* (Alonso-Fernández, Pérez-Colado, et al., 2019; Pérez-Colato et al., 2019), a tool to simplify experiments to validate serious games and/or assess their players. The game sends questionnaires responses as well as in-game interaction data to *Simva* for their storage. Researchers can then access all collected data, conveniently linked by pseudonymous identifiers assigned to each participant, for further analysis.

As well as the responses to both questionnaires, in-game interaction data (traces) were collected during the experiments, including for example interactions with game characters and objects, and general progress within the game's fictional first week at school. All traces were represented using the xAPI-SG Profile. A tracker component embedded into the game prepared these traces as xAPI-SG statements, sending them to an external server, which, in our case, was integrated with *Simva*.

The Experience API (xAPI) (ADL Initiative, 2016) is a format to capture data from e-learning environments based on activity streams (Snell et al., 2011), and was created by a community lead by ADL. Each xAPI trace, also called a *statement*, represents an in-game interaction. Statements in xAPI are

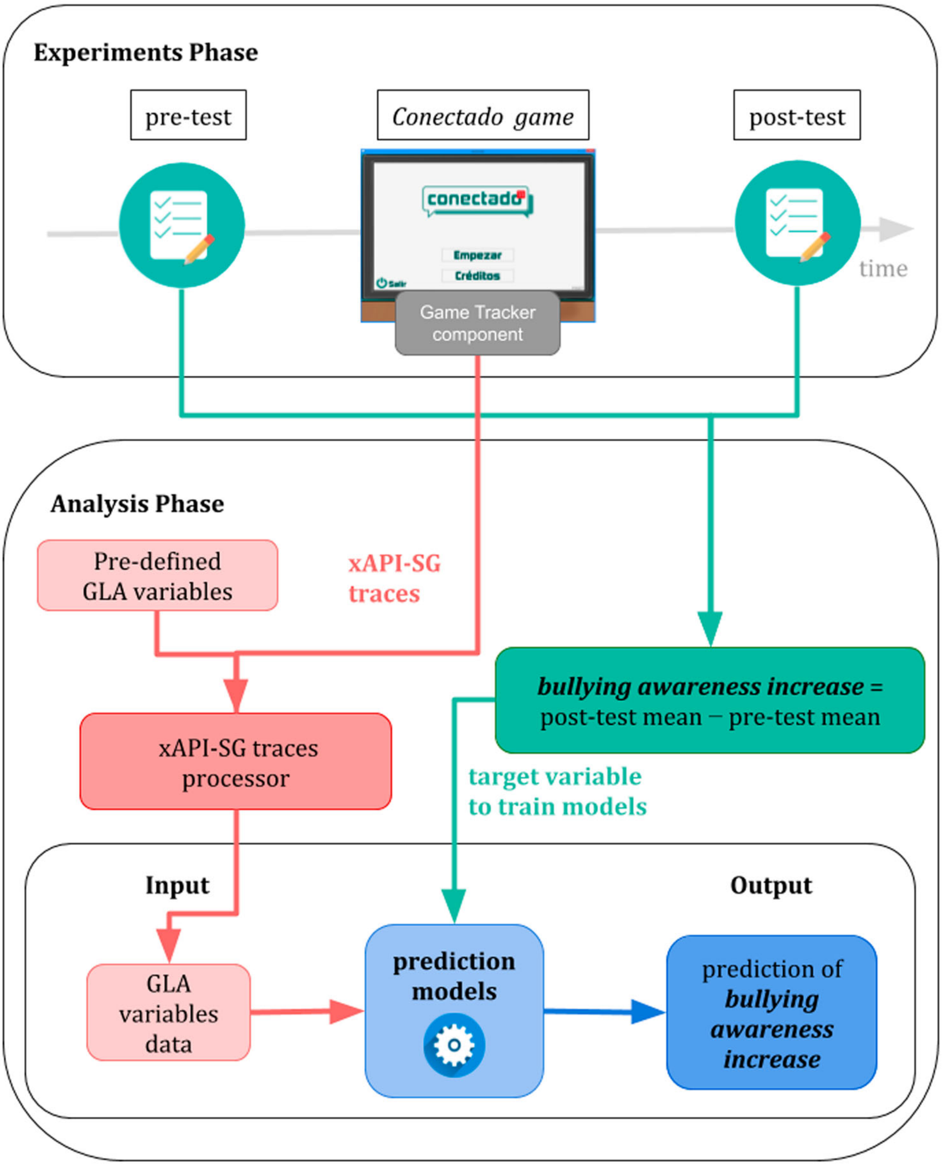


**Figure 1.** Methodology: (1) During the experiments phase, pre-post questionnaires and interaction data from *Conectado* are collected (top); (2) The analysis phase starts with the xAPI-SG traces, from which GLA variables are derived. Then, the GLA variables are used as input for prediction models of the bullying awareness increase.

formatted as JSON and include three main fields – an actor, a verb and an object – and may include additional ones, such as timestamps or results. Domain-specific profiles can be created to use xAPI in specific contexts and/or communities of practice. For serious games, the xAPI-SG Profile (Serrano-Laguna et al., 2017) was created by the e-UCM Research Group in collaboration with ADL, and includes specific verbs that refer to common structural and design elements found in serious games.

The xAPI-SG traces collected from *Conectado* include, as actor, a pseudonymous identifier provided to each player. The verbs used describe relevant in-game interactions: for instance, "initialized" and "completed" are used, respectively, to track the start and end of in-game days. Figure 2 shows an example xAPI-SG trace from *Conectado*, representing an interaction of the player whose identifier

```
{
  "actor": {
    "name": "user-identifier"
  },
  "verb": {
    "id": "http://adlnet.gov/expapi/verbs/interacted"
  },
  "object": {
    "id": "url-of-game-website/game-version/ComputerOnDesk",
    "definition": {
      "type":
      "https://w3id.org/xapi/seriousgames/activity-types/game-object",
    }
  },
  "result": {
    "extensions": {
      "GameDay": 1.0,
      "GameHour": "21:30",
      "MobileMessages": "True"
    }
  },
  "timestamp": "2019-05-17T12:04:56.835Z"
}
```


**Figure 2.** xAPI-SG example statement (trace) tracked from an interaction in *Conectado* when the player uses the in-game computer. Text in italics would be replaced by actual identifiers.

appears in the "actor" field, with an in-game object identified as "ComputerOnDesk" (a computer in the game that the player can interact with). The timestamp of the action is included, and the results contain additional information such as the in-game day that the trace belongs to.

The xAPI-SG statements are then processed to derive the GLA variables to be used in the analysis. For each type of statement, we store the following information:

- For "accessed" actions, an identifier for the target, such as "school_bathroom"
- For "initialized" actions, an identifier for the object of the action, such as the full game or a specific in-game day; and a timestamp.
- For "completed" actions, similar information to that of "initialized"; and, if the full game has been finished, we also store the specific ending reached within the result field.
- For "interacted" actions, the target, which may be an in-game element, for example when using items; or, as is the case for conversations, the character that the player conversed with.
- For "progressed" actions, an object identifier. For example, when tracking the changes in variables that represent the level of friendship with other characters, the identifiers of those characters are used.
- For "selected" actions, the object and the results of the action to track in-game decisions. For example, when players can choose to mention the ongoing bullying to parents, the results would include the player's choice, and the object would identify the point where that choice was taken.

With this information, we can calculate the values of the GLA variables described in the following section.

### *Identifying game learning analytics variables*

Once that all the relevant in-game interactions have been collected as xAPI traces, they can be processed to extract GLA variables in a process known as feature extraction. Although this process is not straightforward, and to a significant extent relies on having in-depth knowledge of the serious game and the exact traces that it generates, we can provide several guidelines and recommendations based on our experience and previous work regarding use of GLA (Alonso-Fernández, Cano, et al., 2019). First, if available, the feature extraction process should be based on the games' Learning Analytics Model (LAM) (Pérez-Colado et al., 2018). The LAM provides the relevant information (events) to be tracked from the game, relating that data with the specifications of the (educational) game design, and describing how that data should be analyzed and/or interpreted. This process could be also guided by a learning analytics reference model (Chatti et al., 2012). If a LAM is not available, we recommend the use of default analysis and visualizations (Alonso-Fernández et al., 2017) based on the main fields of the xAPI-SG Profile, which expose the game's structure and can quickly identify important features to be considered for analysis. Another option is to consult the game designers, who may suggest key elements to analyze from the game.

In *Conectado*, the possible interactions include making choices in the dialogs with characters, such as confronting the bully (Alejandro), mocking a previously bullied character (Maria), or sharing your phone password with bystanders who have also started to bully you. The player can also choose to mention ongoing bullying with the main character's parents or the teacher. After each of these interactions, the level of friendship with other in-game characters can increase or decrease, acting as an indicator of the player's standing regarding those characters. The player can also move within the game's scenarios and interact with objects; for instance, entering the school bathroom when asked to do so by a classmate, attempting to remove chewing-gum placed on clothes by a bully, or using different simulated devices and their (cyberbullying relevant) social applications, such as instant-messaging or social-network apps.

Based on *Conectado*'s LAM, we extracted a set of relevant GLA variables for this game. Some of these variables, such as durations or player-entity interactions, could be obtained directly from xAPI-SG Profile statements, without need for further game-specific knowledge. We additionally included variables representing the in-game actions related to bullying and cyberbullying. From the set of xAPI-SG traces collected from students' gameplays, we extracted the information to fill these pre-defined variables to later use them as input for the prediction models. As an example, variables with the duration of each in-game day for a player are obtained by subtracting the timestamps of the pair of xAPI-SG traces that mention each day as being "initialized" and "completed". Binary variables are extracted by looking up traces that mention their corresponding object identifiers, and then using the results of those traces to arrive at a true or false value. Values of discrete variables are progressively increased as "interaction" traces with the corresponding targets appear.

The 44 GLA variables that we chose to include in our analysis (derived from the xAPI-SG statements) are described in Table 1, including their names, types and brief descriptions. Related variables are described together in the same row of Table 1; for instance, the 5 variables containing the duration of each game day are described together as "duration_day_d, *d* in *[1,2,3,4,5]*". The resulting GLA variables were then used as input for the prediction models described in the next section.

## *Prediction models of bullying awareness increase*

Our final goal is to use the gameplay traces to predict the increase in bullying awareness as a result of playing the game. We define the bullying awareness increase as the difference between the post-test mean score and the pre-test mean score for each player. Therefore, this continuous variable is the target variable for prediction models.

We have used different prediction models to predict the exact value of the increase in bullying awareness and compared predicted results with those obtained in the pre-post-tests. As prediction models, we chose: linear regression, regression trees, Bayesian regression, Support Vector Machines for Regression (SVR), k-nearest neighbors (k-NN), neural networks, random forests, AdaBoost, and gradient boosting. All models were tested with 10-fold cross validation. For all models, different parameters were tuned to find the best ones.

**Table 1.** Variables derived from the xAPI -SG statements of *Conectado*.

| Variable name | Type | Description |
|---|---|---|
| accepted_c, *c* in *[Alison, Guillermo, Jose]* | true/false | Player has accepted a friendship request on in-game computer of character *c* |
| accessed_bathroom | true/false | Player has accessed the school bathroom |
| confront_Alejandro | true/false | Player has confronted Alejandro |
| duration | continuous | Total time playing *Conectado* (in minutes) |
| duration_day_d, *d* in *[1,2,3,4,5]* | continuous | Total time playing day *d* of *Conectado* (in minutes) |
| ending_number | categorical | Ending reached by the player: 1 for worst ending, 2 for regular, and 3 for best ending |
| find_earring | true/false | Player has helped Alison to find her earring |
| friendship_decrease_c, *c* in *[Alejandro, Alison, Ana, Guillermo, Jose, Maria, Parents]* | discrete | Number of times the player has decreased the level of friendship with character *c* |
| friendship_increase_c, *c* in *[Alejandro, Alison, Ana, Guillermo, Jose, Maria, Parents]* | discrete | Number of times the player has increased the level of friendship with character *c* |
| gum_washed | true/false | Player has washed the gum from the clothes |
| has_ended_game | true/false | Player has ended the full *Conectado* game |
| interactions_c, *c* in *[Alejandro, Alison, Ana, Guillermo, Jose, Maria, Mother, Father]* | discrete | Number of interactions the player has carried out with character *c* |
| mock_Maria | true/false | Player has mocked Maria |
| shared_password | true/false | Player has shared the password with classmates |
| tattle_to_parents | true/false | Player has mentioned bullying to parents at home |
| tattle_to_teacher | true/false | Player has mentioned bullying to teacher at the school |
| used_computer | true/false | Player has used the computer at home |
| used_friends_app | true/false | Player has used social network app on smartphone |
| used_mobile_chat | true/false | Player has used instant messaging on the smartphone |

## Results

For each of the 9 prediction models, Table 2 shows the mean absolute error (MAE) and the standard deviation (SD) (normalized to scale [0–10]) for the predictions with the best combination of parameters found for that model.

The model that provides the best results is a Bayesian regression, closely followed by a gradient boosting model, with random forests and Adaboost models at very similar error levels, and all other models providing acceptable results. The difference between the best models is not significant. For the remainder of this section, we focus on the variables that have proven to be most relevant in the best-performing models, which we have identified and attempted to relate to the game design to explain why they appear to have such a great influence on changes in bullying awareness:

- Number of interactions with the character Jose (*interactions_Jose)*: a higher number of interactions predicts higher bullying awareness increase. We consider that a high number of interactions with any character may be a result of a high immersion of the player in the game. This may therefore result in a higher increase in the awareness of the problem.

**Table 2.** Results of predictions of bullying awareness increase.

| Prediction model | Mean Absolute Error (MAE) *normalized to scale [0–10]* | Standard Deviation (SD) *normalized to scale [0–10]* |
|---|---|---|
| Linear regression | 0,581 | 0,047 |
| Regression trees | 0,557 | 0,055 |
| Bayesian ridge regression | 0,540 | 0,053 |
| SVR | 0,556 | 0,051 |
| kNN | 0,578 | 0,048 |
| Neural Networks | 0,557 | 0,050 |
| Random Forests | 0,551 | 0,052 |
| Adaboost | 0,551 | 0,057 |
| Gradient boosting | 0,548 | 0,052 |

- Ending reached (*ending_number)*: a better ending in the game (higher value of ending number) predicts higher increase in awareness. The ending reached is the result of the in-game actions and decisions taken, therefore, it relates to players' behavior in the game. An adequate behavior in a bullying and cyberbullying situation shows a higher awareness of players, either from previous training or from the gameplay, which may show a higher inclination to be attentive in the game and therefore further increase their awareness.
- Duration of in-game day 4 (*duration_day_4*): a higher duration predicts higher increase in awareness. We hypothesize that the specific content of that day may be more relevant for the increase in awareness. Revising its content, it includes the threat to change the player's password, a stolen smartphone, and a case of identity theft. These issues may be more important for players in the target group (12–17 years old), and spending more time on this day, and therefore experiencing them in greater detail, may have a greater impact when increasing awareness.
- Duration of in-game day 3 (*duration_day_3*): in contrast to the previous, duration in day 3 predicts lower awareness increase. We consider that above-average durations may show players losing attention in the game. Revising its content, we notice that the social media has a higher presence, as the main character sees an edited picture of him/herself with classmates making fun of it on the comments. The overlong time on this day may reflect players being distracted by the in-game social media application (which is used as a scripted story element, with minimal actual functionality); and trying to carry out more actions in this app, such as replying to comments may have detracted from gameplay as a whole.

It is important to notice that the information on the prediction relevance of variables, and therefore their discussion and explanation, has only been possible because some of the chosen prediction models allowed the relevance of the input variables to be queried. This would not have been possible for other black-box models, which raises the widely-discussed need of explainable artificial intelligence (xAI), as discussed in (Adadi & Berrada, 2018; Carvalho et al., 2019).

## Discussion

This case-study has showcased the process of performing an evidence-based evaluation of the effect of a serious game, based on in-game user interaction data. The use of the xAPI-SG Profile standard to collect traces proved to be very useful to simplify the evaluation, as it not only simplified the process of collecting and storing traces, but also that of processing and analyzing data. The xAPI-SG is a standard yet powerful and simple format that allowed easy extraction of the information that we were interested in.

We consider that the approach used with *Conectado* can be generalized to other serious games or, at least, to other linear, narrative serious games. The case study builds upon previous work (Alonso-Fernández et al., 2020), in which, using a serious game to teach first aid techniques, the goal was to predict players' final knowledge about the topics covered in the game. In this experience, interaction data was collected and analyzed to derive relevant variables containing information to use as input to predict players' knowledge.

From the common points encountered on the current and previous experience, we consider that this process could be generalized to carry out other evidence-based evaluations of the effectiveness of serious games. The steps followed can be generalized, first, using a standard to track in-game interactions such as the xAPI-SG Profile. Once interaction data are collected, a further step towards generalization is to gather an initial set of variables to derive from the xAPI-SG traces, based on available fields such as the duration of in-game activities, and interactions with relevant in-game items and characters; which can later be complemented with game-dependent information. The initial set of variables can be used as a baseline of what game learning analytics can conclude for the serious game and can be extracted automatically if analytics traces are formatted using an xAPI-SG Profile representation. With those GLA variables, we recommend testing interpretable prediction models that provide information of the relevance of each variable, such as tree-based models (random forests, gradient boosting), which can help to interpret and inform the evaluation process and its results. Moreover, using xAPI allow SGs' developers and researchers to build and reuse a tooling ecosystem for both statements gathering, analysis and predictions.

There are some limitations to the generalizability of this study. First, the fact that the videogame has a narrative, almost-linear structure and a low playing time restricts the variability of the interactions for players. Second, the discussion of the relevance of specific variables in our results is limited by the fact that the prediction model is not a black-box model. Finally, selection of GLA variables is not straightforward and could limit the generalization of our approach; however, we have provided guidelines and recommendations to identify an initial set of GLA variables for use in prediction.

## Conclusions

We have showcased a full example of an evidence-based evaluation of the effectiveness of a serious game from interaction data: from the possible interactions in the game, their collection using the xAPI-SG standard, the analysis of the resulting traces to derive variables, and the use of those variables as input for prediction models, which, once trained (with pre-post experiment data), provide evaluations of increase of bullying awareness in players based solely on in-game interaction data.

The experiments phase has shown the convenience and advantages of using a standard data format to simplify collection, processing and analysis of interaction data. As the xAPI-SG format is well-defined, no further processing is required, and its use greatly simplifies the feature extraction process. The analysis phase has described the process of bridging the gap between in-game interactions (collected as xAPI-SG traces) and relevant information (stored as GLA variables). The feature extraction was performed based on our knowledge from the game, although many variables could have been selected based solely on the fields available in the xAPI-SG Profile, since the use of this standard format also exposes the general structure of games where it is used. The explainability of the results, which in our case was possible due to our choice of prediction models, is especially relevant towards the deployment of our approach in educational scenarios, as it is required to be able to explain how the results have been obtained, both to teachers and to students evaluated with such techniques.

It is also important to notice that the serious game used on this case-study was not originally designed to be an assessment tool; its mechanics and interactions were solely designed to depict a bullying situation which would make players reflect on their actions and their consequences. Some changes in the game could improve the results: a higher variety of options, leading to even more endings, could provide deeper insight on the actions that players will take in similar situations, at the expense of an increase in game complexity. In the same sense, a broader set of interactions with the characters could provide more information about players' attitudes with the different profiles (e.g. bully, observer, person previously bullied), although this would again increase the game's complexity.

Based on the current and previous work, we consider that most of the steps that we followed could be further generalized, using the xAPI-SG standard to capture interaction data and defining a minimum set of variables to extract from xAPI-SG traces, thus establishing a standard, automated process for evidence-based evaluations of the effectiveness of serious games.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work has been partially funded by Regional Government of Madrid (eMadrid S2018/TCS-4307, co-funded by the European Structural Funds FSE and FEDER), by the Ministry of Education (TIN2017-89238-R), by the European Commission (Erasmus+ IMPRESS 2017-1-NL01-KA203-035259) and by the Telefónica-Complutense Chair on Digital Education and Serious Games.

## Notes on contributors

***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.

***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.

## ORCID

*Cristina Alonso-Fernández* http://orcid.org/0000-0003-2965-3104
*Antonio Calvo-Morata* http://orcid.org/0000-0001-8701-7582
*Manuel Freire* http://orcid.org/0000-0003-4596-3823
*Iván Martínez-Ortiz* http://orcid.org/0000-0001-6595-5690
*Baltasar Fernández-Manjón* http://orcid.org/0000-0002-8200-6216

## References

Adadi, A., & Berrada, M. (2018). Peeking inside the Black-Box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

ADL Initiative. (2016). *xAPI specification*. 2014. https://github.com/adlnet/xAPI-Spec/blob/a752217060b83a2e15dfab69f8c257cd86a888e6/xAPI.md

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, *141*, 103612. https://doi.org/10.1016/j.compedu.2019.103612

Alonso-Fernandez, C., Calvo, A., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2017, April 25–28). *Systematizing game learning analytics for serious games*. 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, Greece, 1111–1118. IEEE. https://doi.org/10.1109/EDUCON.2017.7942988

Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior*, *99*, 301–309. https://doi.org/10.1016/j.chb.2019.05.036

Alonso-Fernández, C., Pérez-Colado, I. J., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). *Using* Simva *to evaluate serious games and collect game learning analytics data*. LASI Spain 2019: Learning Analytics in Higher Education (pp. 22–34). http://ceur-ws.org/Vol-2415/paper03.pdf.

Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., & Fernández-Manjón, B. (2020). Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *Journal of Computer Assisted Learning*, *36*(3), 350–358. https://doi.org/10.1111/jcal.12405

Álvarez García, D., Núñez Pérez, J., & Dobarro, A. (2013). Cuestionarios para evaluar la violencia escolar en Educación Primaria y en Educación Secundaria: CUVE3-EP y CUVE3-ESO. *Apuntes de Psicología*, *31*(2), 191–202.

Calderón, A., & Ruiz, M. (2015). A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education*, *87*, 396–422. https://doi.org/10.1016/j.compedu.2015.07.011

Calvo-Morata, A., Freire-Moran, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2019). Applicability of a cyberbullying videogame as a teacher tool: Comparing teachers and educational sciences students. *IEEE Access*, *7*, 55841–55850. https://doi.org/10.1109/ACCESS.2019.2913573

Calvo-Morata, Antonio, Rotaru, Dan Cristian, Alonso-Fernandez, C., Freire-Moran, M., Martinez-Ortiz, I., & Fernandez-Manjon, B.. (2020). Validation of a cyberbullying serious game using game analytics. *IEEE Transactions on Learning Technologies*, *13*(1), 186–197. https://doi.org/10.1109/TLT.2018.2879354

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832. https://doi.org/10.3390/electronics8080832

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5/6), 318. https://doi.org/10.1504/IJTEL.2012.051815

Freire, M., Serrano-Laguna, Á, Iglesias, B. M., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game learning analytics: Learning analytics for serious games. In M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology* (pp. 1–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_21-1

Garaigordobil, M., & Aliri, J. (2013). Ciberacoso ("cyberbullying") en el País Vasco: Diferencias de sexo en víctimas, agresores y observadores. *Behavioral Psychology/ Psicologia Conductual*, *21*(3), 461–474.

Kato, P. M., & Klerk, S. D. (2017). Serious games for assessment: Welcome to the jungle. *Journal of Applied Testing Technology*, *18*, 1–6.

Liu, M., Kang, J., Liu, S., Zou, W., & Hodson, J. (2017). Learning analytics as an assessment tool in serious games: A review of literature. In M. Ma & A. Oikonomou (Eds.), *Serious games and edutainment applications* (pp. 537–563). Springer International Publishing. https://doi.org/10.1007/978-3-319-51645-5_24

Ortega-Ruiz, R., Del Rey, R., & Casas, J. A. (2016). Evaluar el bullying y el cyberbullying validación española del EBIP-Q y del ECIP-Q. *Psicología Educativa*, *22*(1), 71–79. https://doi.org/10.1016/j.pse.2016.01.004

Peng, W., Lee, M., & Heeter, C. (2010). The effects of a serious game on role-taking and willingness to help. *Journal of Communication*, *60*(4), 723–742. https://doi.org/10.1111/j.1460-2466.2010.01511.x

Perez-Colado, I., Alonso-Fernandez, C., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2018, April 17–20). *Game learning analytics is not informagic!*. 2018 IEEE Global Engineering Education Conference (EDUCON), Tenerife, Spain, 1729–1737. IEEE. https://doi.org/10.1109/EDUCON.2018.8363443

Perez-Colado, I. J., Calvo-Morata, A., Alonso-Fernandez, C., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2019, July 15–18). *Simva: Simplifying the scientific validation of serious games*. 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), Maceió, Brazil, 113–115. IEEE. https://doi.org/10.1109/ICALT.2019.00033

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces*, *50*, 116–123. https://doi.org/10.1016/j.csi.2016.09.014

Shoukry, L., Göbel, S., & Steinmetz, R. (2014). *Learning analytics and serious games: Trends and considerations*. Proceedings of the 2014 ACM International Workshop on Serious Games. https://doi.org/10.1145/2656719.2656729

Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_4

Shute, V., & Ventura, M. (2013). Stealth assessment. In J. Michael Spector (Ed.), *The SAGE encyclopedia of educational technology* (p. 91). SAGE Publications. https://doi.org/10.4135/9781483346397.n278

Snell, J., Atkins, M., Norris, W., Messina, C., Wilkinson, M., & Dolin, R. (2011). JSON activity streams 1.0. *Search ResultsWeb resultsActivity Streams Work*, *22*(8), 2013.

Xu, Y., Johnson, P. M., Lee, G. E., Moore, C. A., & Brewer, R. S. (2014). Makahiki: An open source serious game framework for sustainability education and conservation. *International Association for Development of the Information Society*, *8*, 131–138.