Received: 5 June 2019 | Revised: 14 October 2019 | Accepted: 3 November 2019

DOI: 10.1111/jcal.12405

ARTICLE

Journal of Computer Assisted Learning WILEY

Predicting students' knowledge after playing a serious game based on learning analytics data: A case study

Cristina Alonso-Fernández | Iván Martínez-Ortiz | Rafael Caballero | Manuel Freire | Baltasar Fernández-Manjón

Computer Science Faculty, Complutense University of Madrid, Madrid, Spain

Correspondence
Cristina Alonso-Fernández, Computer Science Faculty, Complutense University of Madrid, Madrid, Spain.
Email: calonsofernandez@ucm.es

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12405.

Abstract

Serious games have proven to be a powerful tool in education to engage, motivate, and help students learn. However, the change in student knowledge after playing games is usually measured with traditional (paper) prequestionnaires–postquestionnaires. We propose a combination of game learning analytics and data mining techniques to predict knowledge change based on in-game student interactions. We have tested this approach in a case study for which we have conducted preexperiments–postexperiments with 227 students playing a previously validated serious game on first aid techniques. We collected student interaction data while students played, using a game learning analytics infrastructure and the standard data format Experience API for Serious Games. After data collection, we developed and tested prediction models to determine whether knowledge, given as posttest results, can be accurately predicted. Additionally, we compared models both with and without pretest information to determine the importance of previous knowledge when predicting postgame knowledge. The high accuracy of the obtained prediction models suggests that serious games can be used not only to teach but also to measure knowledge acquisition after playing. This will simplify serious games application for educational settings and especially in the classroom easing teachers' evaluation tasks.

KEYWORDS

learning analytics, serious games, game-based learning, assessment, e-learning, xAPI

1 | INTRODUCTION

Serious games (SGs) are games or game-like applications with purposes beyond entertainment (Michael & Chen, 2005). In education, SGs have proven to be an effective way to promote learning due to their engaging and immersive nature (Boyle, Connolly, Hainey, & Boyle, 2012), which increases students' participation in the learning process.

To adequately evaluate players' knowledge when using an SG, a common method is the use of two external questionnaires for each player, one *before* playing (pretest) and another *after* playing (posttest). This methodology is the most common and accepted practice in the medical domain to evaluate the efficacy of SGs (Calderón & Ruiz, 2015). However, several authors have pointed out that this external and summative evaluation of learning is error prone and reduces the time to play (Clark, Martinez-Garza, Biswas, Luecht, & Sengupta, 2012; Frederick-Recascino, Liu, Doherty, Kring, & Liskey, 2013). Preexperiments–postexperiments can also be used to measure how students' knowledge improves when using an already evaluated game. This evaluation of acquired knowledge is the focus of this work. The goal of this research is to showcase the use of in-game interactions to predict students' knowledge using data mining techniques.

The goal of learning analytics (LA) techniques is to collect, analyse and report data to understand and optimize learners' contexts on educational systems (Long & Siemens, 2011). The high interactivity of SGs provides large quantities of interaction data, allowing the application of LA techniques. Game learning analytics (GLA) is defined as the

process of capturing, storing, analysing, and obtaining information from players' interactions with an SG (Freire et al., 2016).

We consider that in-game interactions (i.e., GLA) can be analysed to automatically and accurately determine users' knowledge after playing. This allows us to evaluate players as an integral part of the playing, avoiding disruption of the game experience and without needing an external measure. We propose to determine players' knowledge following these stages:

- Game validation phase: The SG is validated using the traditional and widely accepted prequestionnaires–postquestionnaires, while also tracking players' interactions. Then, different supervised machine learning models are tested to predict knowledge, taking as input the interaction data (and, optionally, the pretest) and validated against actual knowledge results (given in the posttest).
- Game deployment phase: Once a sufficiently accurate prediction model is obtained, in subsequent applications of the game, students' knowledge after playing can be automatically predicted on the basis of in-game interactions. This greatly simplifies the application of games in the classroom by no longer requiring students to fill prequestionnaires–postquestionnaires. The predicted students' knowledge can be used as an indicator for teachers to know how much students finally know about the topic covered in the game.

We test our approach by conducting a case study to determine if players' knowledge, as measured by a posttest, could be accurately predicted by applying machine learning techniques to previously gathered information (pretest and in-game interactions). We are also interested in determining the best prediction models and the most relevant information when predicting knowledge. Additionally, we want to know the extent to which availability of the pretest (which directly measures players' knowledge before the game) affects the accuracy of predictions.

The rest of the paper is structured as follows: Section 2 reviews the related work for data mining techniques applied for knowledge predictions in education; Section 3 states the research questions of the current case study; Section 4 describes the methodology, including participants, experimental design, and materials and instruments; Section 5 summarizes the best results obtained from the predictions models; Section 6 presents a discussion of the results in relation to the research questions; and Section 7 contains conclusions, limitations, and future work.

2 | RELATED WORK

Data mining techniques have been applied in education to understand students and their learning scenarios, in a discipline called educational data mining (EDM; Baker & Yacef, 2009). For instance, the U.S. Department of Education has studied the use of data mining techniques for different purposes such as prediction to enhance learning (Bienkowski, Feng, & Means, 2012). On page 28, they highlight that "inferring what a user knows […] requires looking at accumulated data that represents the interactions between students and the learning system." Data mining techniques have been applied to LA data to predict students' knowledge and prevent their failure, helping teachers and students to improve their teaching and learning processes (Shahiri, Husain, & Rashid, 2015). These predictions usually target performance, knowledge, score, or marks, either via regression analysis to find relationships between students' variables or via classification to group students (Romero & Ventura, 2010). Authors have also revised the models used to predict students' knowledge finding that they commonly include neural networks and decision trees (Shahiri et al., 2015), Bayesian networks, rule-based systems, regression, and correlation analysis (Romero & Ventura, 2010). The analysis of 240 EDM publications by (Peña-Ayala, 2014) yielded that student modelling and assessment are the main targets of EDM application, with predictive models for classification being the most frequently applied, in particular Bayes theorem, logistic regression, and decision trees. Support vector machine models have also been used to predict faculty performance evaluation, using different kernel methods (Deepak, Pooja, Jyothi, Kumar, & Kishore, 2016). On a recent literature review, we found out that assessment is commonly the target of the application of data mining techniques to LA data, specially applying classical techniques such as regression and decision trees (Alonso-Fernández, Calvo-Morata, Freire, Martinez-Ortiz, & Fernández-Manjón, 2019).

Despite the highly accurate results obtained by applying data mining to LA data, we have not found widely accepted approaches that systematically measure players' knowledge after using SGs. The instruments that measure knowledge (usually prequestionnaires–postquestionnaires) have to be developed ad hoc for each game (Petri & Gresse von Wangenheim, 2016). This is a significant limitation that influences a lack of replicability and of well-defined models (Petri & Gresse von Wangenheim, 2017). In the last years, data-based evaluations have been used to measure the learning process in a discipline called *stealth assessment* (V. Shute & Kim, 2014). Stealth assessment relies on capturing sequences of actions made by students while interacting with a highly interactive and immersive tool (e.g., a game), to obtain information of what students know and do not know at each moment (V. Shute & Ventura, 2013). This information is updated when new data are captured and is later used to evaluate students. This promising research line has high implementation cost yet, as solutions need to be developed ad hoc for each game. We consider that our two-step approach can be more easily generalized to a broader set of educational games. To the best of our knowledge, we have not found specific studies where data mining has been used to improve the evaluation of students' knowledge when using SGs that had already been experimentally validated, as we propose in our approach.

3 | RESEARCH QUESTIONS

To test our two-step approach, we must first verify that we can infer students' knowledge after playing an SG. This motivates the initial research question of this case study:

- Q1.1. Can we accurately predict student knowledge from previous knowledge and interactions with an SG? (we refer to this as the *pre + game* condition, because we use both pretest and in-game interactions to build the prediction models).
- In case we can predict it, our next step is to find the most accurate predictions models and the most relevant information for those predictions. Therefore, we also propose a follow-up research question:
- Q1.2. If we can indeed predict student knowledge after playing an SG, what prediction models perform best, and what are the most relevant variables for these models?

We also want to explore the possibility of predicting knowledge solely from in-game interactions, without relying on any pretest information. We call this the *game-only* condition. For this condition, we again look for the most suitable models and the most relevant information for predictions and will compare it against results from Q1.1 and Q1.2, which use the *pre + game* condition. Therefore, we propose the following additional research questions:

- Q2.1. Can we accurately predict student knowledge solely from interactions with an SG? (*game-only* condition)
- Q2.2. What are the best prediction models and the most valuable information towards those predictions?
- Q2.3. Is the *pre + game* condition (proposed in Q1.1) more effective at predicting student knowledge than the *game-only* condition (proposed in Q2.1)?

To answer these research questions, we have used a previously validated SG on a preexperiment–postexperiment with a control group (Marchiori et al., 2012). As the game was later updated to a new technology, to carry out the "Game validation phase" of our approach and build the prediction models, we first conducted a new set of experiments with prequestionnaires–postquestionnaires while tracking GLA data from players' interactions.

4 | METHODOLOGY

4.1 | Participants

The experiments for this case study involved N = 227 high school students from a charter school in Madrid, Spain. We conducted two sessions with 28 students as an initial formative evaluation. Their feedback helped us to test the remote data collection in the school settings and prepare for the main experience. Out of the remaining population (N = 199), gender was not obtained for 15 students due to an error handling a questionnaire. For the other 184 students, the gender distribution was 46.7% males and 53.3% females. The median age was 14 years old. Figure 1 summarizes the gender distribution by age. In terms of gender and age, this sample is representative of the student population in Madrid (Comunidad de Madrid, 2016; Instituto Nacional de Evaluación Educativa—Ministerio de Educación, 2017).

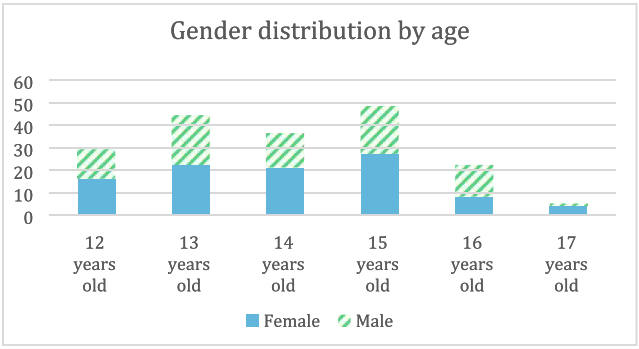

FIGURE 1 Gender distribution by age of the participants in the experiments

4.2 | Experimental design

At the beginning of each session, the teacher (playing the role of the session manager) gave each student a unique identification code that allowed them to access the game and which was used in all questionnaires instead of any personally identifying information (Perez-Colado et al., 2019). Then, each student/player completed, in this order: (a) a questionnaire before starting the game (pretest), (b) a complete game session in the chosen SG, and (c) a questionnaire after playing the game (posttest). Each player is therefore linked to the three data sources via the unique player identification code, which acts as a pseudonym and reduces potential privacy pitfalls. The complete experiment was designed to fit into a standard 50-min session. Students could repeat the game levels as many times as they wanted up to 30 min. Figure 2 summarizes the research design of the experiment. The experiment was reviewed and approved by the school management as an educational activity. Students were informed about the data capturing, and the school signed an informed consent.

4.3 | Materials and instruments

4.3.1 | The *First-Aid Game*

The *First-Aid Game* is a game-like simulation with narrative structure that aims to teach basic life-support manoeuvres for players for 12–16 years old, focusing on chest pain, unconsciousness, and

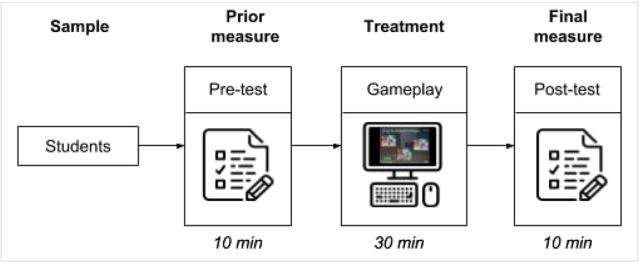

FIGURE 2 Research design of the experiment: All players completed a pretest, a game session, and a posttest

FIGURE 3 Textual and visual options available in the serious game *First-Aid Game*

choking. These three situations are depicted as game levels. In each scenario, players can interact with the main character or use a mobile phone (visible at the bottom right corner in the first screenshot of Figure 3) to call the emergency services. The game offers multiple-choice situations (second screenshot of Figure 3) that feature the specific first aid knowledge to be learnt through the game (e.g., Heimlich manoeuvre to avoid chocking). Players learn whether their decisions are appropriate or not: If they choose an incorrect answer, either the game reports the critical error and its consequences and lets them try again until they choose the correct answer or the game allows you to continue to later discover the consequences (and it is reflected in the final score). Options may be textual or visual (see Figure 3). The game includes random elements to improve reflection and replayability (e.g., availability of a semi-automatic external defibrillator). The three levels can be replayed as many times as desired during the available time. After each level is completed, a score is provided to indicate players whether their actions were mostly correct or not. This score does not directly measure players' knowledge but challenges them to replay levels where they made many mistakes.

The game was developed and evaluated by the e-UCM Research Group and actual emergency physicians (e-UCM, 2012; Marchiori et al., 2012). The game was validated in 2012 with an experiment that included pretests–posttests to measure players' knowledge and a control group to compare the game effect against that of a theoretical–practical demonstration by a trained instructor. Players in the experimental group gained, on average, 2.07 points on a 10-point scale, compared with control group learners who gained 3.61 points. This proved that the game achieves its goal of making player learn first aid procedures. The game was later adapted and updated to the Unity 3D videogame engine using *uAdventure* (Perez Colado, Perez Colado, Martínez-Ortiz, Freire, & Fernandez-Manjon, 2017), an authoring tool developed by the same group. This included the tracking of GLA data used in the present work.

4.3.2 | Questionnaires

As mentioned above, two questionnaires were used in the experiments. The pretest consisted of three parts: demographic variables (players' gender and age); a first aid knowledge questionnaire with 15 multiple-choice questions, also used in the original experiment to validate the game (Marchiori et al., 2012) and covering the game contents; and a game habits questionnaire with eleven 5-point Likert questions on game habits obtained from (Manero, Torrente, Freire, & Fernández-Manjón, 2016) and slightly adapted for this experiment. The posttest consisted of two parts: a repetition of the first aid knowledge questionnaire used in the pretest (to compare results) and a questionnaire to evaluate the experience itself, with five 5-point Likert questions assessing the experience, and optional free-text sections for feedback. The scores on the first aid knowledge questionnaires are defined as the total number of correct answers. Therefore, possible scores ranged from 0 to 15 points. Internal consistency of the scale used was ensured when the test was created in the original validation experiment (Marchiori et al., 2012). These questionnaires were a simplification of the ones used in the medical domain and had been previously validated.

4.3.3 | GLA data collected

During gameplays, a software component embedded in the SG (called a *tracker*) sent out players' interactions (i.e., traces) to an external server, developed by the e-UCM Research Group, using the Experience API (xAPI) standard's SG profile (Serrano-Laguna et al., 2017) to transmit and store interaction data.

The collected xAPI data[1] were analysed to derive variables that described how each player played the game. The specific information to be tracked from the game as well as the derived variables were chosen on the basis of the learning and game designs of the game as specified in its LA model (Perez-Colado, Alonso-Fernández, Freire-Moran, Martinez-Ortiz, & Fernández-Manjón, 2018) and in collaboration with domain experts. The process of capturing the data following the LA model for this and two other games is described in more detail in Alonso-Fernández et al. (2019). The variables extracted from the in-game interactions included whether the game has been completed or not, the first and maximum scores achieved in each of the three game levels, the number of times each level was repeated, the interactions

[1]The data that support the findings of this study are available from the corresponding author upon reasonable request.

with game elements, and whether specific questions were answered correctly or not. Appendix A provides the full list of variables derived from the xAPI statements and their detailed descriptions.

4.3.4 | Prediction models

All prediction models were built using RStudio and taking, as inputs, the complete set of variables derived from xAPI data, as described above. Models were created with and without pretest information as input, to further determine if the pretest is essential to predict players' knowledge after playing or not. The target variable of the predictions is the posttest score. Two types of models were created: linear models to predict exact score in range (0–15), and classification models to predict pass/fail category (establishing pass as 8 points out of 15).

We selected the algorithms most widely used in the literature for data mining applied to LA data: regression and decision trees, and linear and logistic regression. Although trees can show complex, non-linear relationships providing easy-to-understand models, regression is useful when data are not extremely complex or not a lot of data are gathered. Additionally, these models are white box models, which will allow us to relate the results obtained to our input data to obtain further information related to the traces collected from the game. A priori, our dataset is not too large, so regression should still be viable; however, if complex relationships appear, trees are expected to be better at discovering them. Different models were tested, including and excluding variables and interactions between variables. We additionally included two methods commonly mentioned in the literature: naïve Bayes for classification and support vector machines for regression (SVR), testing different non-linear kernels (polynomial, radial basis, and sigmoid; Drucker, Burges, Kaufman, Smola, & Vapnik, 1997), and tuning the different parameters, with the ranges recommended in the literature (Hsu, Chang, & Lin, 2016). Models were compared using 10-fold cross-validation. When predicting pass/fail, and because data were not balanced (169 students passed the posttest, whereas only 30 failed it), classification models were created with an undersample of 78 students (40% from the fail class and 60% from the pass class) and tested on the original sample.

5 | RESULTS

We first verified that again in this case, study knowledge increase was significant. Pretest and posttest score variables were not normal (Shapiro–Wilk test yielded $p < .01$). Therefore, to measure knowledge change without assuming a normally distributed population, we use the paired sample Wilcoxon signed-rank test. The test showed a significant increase ($p < .05$, $r = -41$) from pretest scores (mean = 8.06, SD = 2.05) to posttest scores (mean = 9.83, SD = 2.38). This proves replicability of results from the validation experiment and allows us to create predictive models.

As stated in the previous section, we used decision trees, logistic regression, and naïve Bayes classifier for pass/fail predictions and regression trees, linear regression, and SVR with non-linear kernels for score predictions. For pass/fail predictions, Table 1 provides precision, recall, and error, measured as misclassification rate. For score predictions, we provide the mean and the standard deviation of the error. Notice that the error is measured in the score scale of 0–15. Table 1 also summarizes the best models obtained, highlighting the best results in bold font.

The first rows of Table 1 summarize the best models when predicting pass/fail and posttest score with both pretest and in-game information (*pre + game* condition): Logistic regression provides the lowest misclassification rate and the highest recall when predicting pass/fail, whereas SVR provides the lowest mean error for score predictions (although the standard deviation is higher than for regression trees). The lower half of the table is dedicated to the *game-only* condition, summarizing results from models that predict pass/fail and posttest scores solely with in-game information. Logistic regression again provides the most accurate predictions of pass/fail, whereas SVR methods provide the lowest mean error when predicting score.

For score prediction models, 95% confidence intervals (CIs) for predictions were calculated using bootstrapping. The score scale of 0–15 was used, and then results were normalized to the 0–10 scale typically used for grading in Spain. In the *pre + game* condition, the regression tree obtained a mean posttest score prediction (in 0–10 scale) of 6.56 with 95% CI of [3.74, 8.53], whereas linear regression obtained a mean prediction of 6.62 with 95% CI of [4.76, 7.53]. In the *game-only* condition, the regression tree obtained a mean score

TABLE 1 Prediction models for posttest pass/fail and score, with and without pretest information

	Pass/fail prediction				Score prediction (scale [0–15])	
		Success measure		Error		
Pretest?	Data mining model	Precision	Recall	MR	Data mining model	Mean (*SD*)
Yes (*pre + game*)	Decision tree	81.6%	94.2%	16.2%	Regression tree	2.22 **(0.55)**
	Logistic regression	89.8%	**98.3%**	**10.5%**	Linear regression	1.68 (1.44)
	Naïve Bayes classifier	**92.6%**	89.7%	15.1%	SVR (non-linear kernels)	**1.47** (1.33)
No (*game-only*)	Decision tree	88.6%	92.4%	17.3%	Regression tree	2.38 **(0.62)**
	Logistic regression	87.2%	**98.8%**	**12.7%**	Linear regression	1.89 (1.54)
	Naïve Bayes classifier	**89.7%**	90.6%	16.9%	SVR (non-linear kernels)	**1.56** (1.37)

prediction of 6.54 with 95% CI of [4.06, 8.42], whereas linear regression obtained 6.55 with 95% CI of [4.6, 7.74]. These CI results confirm that linear regression models outperform regression tree models in terms of accuracy for score prediction.

Not all the variables used as input for the models (listed in Appendix A) were found to have the same relevance towards predictions. With pretest information, in predicting pass/fail results, the most relevant variables were the pretest score, the final game score, and the number of times each situation was repeated. In predicting score, the most relevant were the maximum score achieved in the "chest pain" game level, the number of interactions with the game character, and failure when answering the question regarding the Heimlich position. Solely with game interactions, most important variables to predict pass/fail include interactions with the game character and the first and maximum score achieved in the "chest pain" level. To predict score, the number of interactions with the game character appears as a relevant variable in all models, together with the maximum score in both the "chest pain" and "unconsciousness" levels, and failure on the "Heimlich position" question.

6 | DISCUSSION

In this section, we answer the research questions stated in Section 3, on the basis of discussing the results presented in Section 5.

Q1.1. Can we accurately predict student knowledge from previous knowledge and interactions with an SG? (*pre + game* condition)

Yes. The highly accurate results allow prediction of knowledge (as posttest results) from previous information (pretest and game interactions). As expected, more accurate predictions are obtained for pass/fail categories, but score predictions are still reasonably accurate. In many situations, as this case study not dealing with core subjects (e.g., math), it may suffice for teachers to know if students have acquired enough knowledge to pass or fail the subject. In fact, classification is most widely used for education (Peña-Ayala, 2014; Shahiri et al., 2015).

Q1.2. If we can indeed predict student knowledge after playing an SG, what prediction models perform best, and what are the most relevant variables for these models?

To classify players in pass/fail categories, the best model is a logistic regression, as naïve Bayes has higher precision but lower recall. This is not surprising, as we are predicting a binary variable, a task well suited for logistic regression (Maalouf, 2011), instead of classifying among several categories. To predict posttest scores, results are not as precise, but SVR yields the lowest error. As SVR was built with non-linear kernels, this result may be an indicator of non-linear relationships between the variables.

The most valuable variables for these predictions include the number of interactions with the game character, the final game score, and the maximum score achieved in the "chest pain" level, although some pretest variables were also slightly significant (e.g., pretest score). A possible explanation is that game mechanics and educational design relate scores in each level (and final score) to knowledge, low scores being a consequence of making domain-relevant errors. Although scores in this specific game are a good indicator of knowledge, this may not be the case for all games. For interactions with the game character, models show that higher number of interactions predicted lower scores. As in most situations, the game design forces players to retry when making an error, and the number of interactions will increase when errors are made, suggesting a "trial and error" strategy. This design decision may explain why the number of interactions is a good predictor of knowledge. If this mechanic appears in other games, a good predictor may be the number of retries or errors. Notice that this discussion is possible as we are analysing results of a white box model (logistic regression); for black box models, such a discussion, if possible, will not be as straightforward (Dreiseitl & Ohno-Machado, 2002).

Q2.1. Can we accurately predict student knowledge solely from interactions with an SG? (*game-only* condition)

Yes. We have obtained accurate prediction results for posttest scores solely from in-game interactions, without pretest information. More accurate results are obtained when predicting pass/fail classification, but still accurate results are obtained for score predictions.

Q2.2. What are the best prediction models and the most valuable information towards those predictions?

To predict pass/fail results, the best prediction model appears to be a logistic regression, as in the case of the *pre + game* condition. Although other models provide a slightly better precision, recall is higher and misclassification rate much lower than in the other models. To predict posttest scores, in the 0–15 range, the best prediction model is again based on SVR. However, we notice that the standard deviation is higher than in the decision tree (which has a higher mean error).

The most useful variables for these predictions again include the number of interactions with the game character and the first and maximum scores obtained in the "chest pain" game level. Regarding interactions with game character, the higher the number of interactions, the lower the score predicted, so the same discussion as above is valid. An unexpected finding is the greater relevance of (first and maximum) scores in the "chest pain" level compared with those of the two other levels. A possible explanation is that, although players could play levels in any order, the "chest pain" level appears in the left-most part of the screen, so most students, accustomed to scanning media left to right (Spalek & Hammad, 2005), played it first. Therefore, this result suggests that the first level students play may have a greater influence on knowledge acquisition.

Q2.3. Is the *pre + game* condition (proposed in Q1.1) more effective at predicting student knowledge than the *game-only* condition (proposed in Q2.1)?

Yes but only slightly. Models in the *pre + game* condition show better predictions in general, but, as shown in Table 1, models in the *game-only* condition still obtain accurate results. That is, results show similarly accurate predictions with and without pretest information, if in-game interactions remain as input for the models.

7 | CONCLUSIONS AND LIMITATIONS

This work presents a case study of an approach to measure students' knowledge after playing SGs based on their game interactions, after an initial priming phase to create the prediction models. We have tested the "Game validation phase" of our approach with an already validated game to ensure that the game indeed fulfils the goal of making players learn. The high accuracy of the models obtained on this case study show that we can indeed predict knowledge after playing, using both pretest and game LA data as inputs. From the models tested, we have seen that pretest information is relevant but by no means essential. Therefore, it is possible to infer students' knowledge solely from in-game interactions. Another option is to use the pretest as a formal evaluation of previous knowledge and, comparing it with the predictions of subsequent knowledge, calculate how much players have learned playing.

After the initial phase to formally validate the game and train the algorithms, games could be deployed automatically (with no posttest required), because the knowledge gained by playing can be inferred from game interactions, as described in the "Game deployment phase" of our approach. Using the most accurate prediction model, a prediction of knowledge after playing the game for each player could be obtained without the need to carry out the posttest. This prediction could be used for teacher as evaluation, allowing the use of games not only to teach but also to measure knowledge gained by players, while reducing costs of experiments in both time and effort. This approach also allows games to be played by larger samples of students, whose results can be automatically predicted. In some scenarios (e.g., in online games), after playing, students could optionally accept the score predicted from their interactions or take a real posttest. Another possibility for teachers is to use the games as an exam to evaluate students, taking the predicted knowledge as their score.

The encouraging results obtained on this case study suggest that our two-step approach proposed may be generalized at least to other similar cases, such as games for procedural learning or game-like simulations with narrative structure that are quite common in several domains (e.g., military and medicine). Both can provide similar interaction data, and therefore, by following the described steps, a similar approach could be applied. The specific data to be collected on each case should be driven by the specific educational design of each game, similarly to what we have done in our case study following the LA model (Perez-Colado et al., 2018), although we expect that relevant interactions for one type of game will also be relevant for similar games.

Based on our results, we can extract some lessons learned, which may be useful for SG designers and researchers that wish to assess players with SGs. For our game, we have found specific data (e.g., number of interactions with game elements) that seem to be related with knowledge and an emphasis on the results achieved on the early phases of the game (e.g., scores in game levels). In Section 6, we have provided possible explanations for these results linked with both the game mechanics and the educational game design. We therefore advise both mechanics and educational designs to be considered when deciding which interaction data to capture from SGs. Using an accepted standard format (e.g., xAPI-SG) is a clear recommendation, as it simplifies tracking, replicability of models, and integration in a wider range of systems.

This work has some limitations. The most relevant is that the SG used was evaluated in a previous preexperiment–postexperiment using an accepted existing measurement test on the topics of the game (Marchiori et al., 2012). This allowed us to apply prediction models, as it was proven that students learned with the game. Other limitation is that the data used are from one SG and a single school, which could potentially bias the results. However, we consider that the approach could be generalized for a wider range of games and students with similarly accurate results.

Future lines of work include testing these approaches with larger datasets and more complex games to attempt to replicate the highly encouraging results reported in this work. We also consider that better results could be obtained using games with knowledge prediction as an explicit design goal. Therefore, another line is testing this approach with games originally designed to be evaluated with these techniques, for instance, designing both the game and the interaction data to be gathered to improve score predictions. The predictions of learning obtained may also be used for players' assessment. We plan to study the use of games for assessment and propose a similar approach as the one described on this work focusing on the in-game stealth assessment of players (V. J. Shute & Moore, 2017). Although the most promising algorithms identified in the literature (Deepak et al., 2016; Peña-Ayala, 2014; Romero, López, Luna, & Ventura, 2013; Romero & Ventura, 2010; Shahiri et al., 2015) have been tested, some other more complex or even nontraditional methods could also be explored.

ORCID

Cristina Alonso-Fernández https://orcid.org/0000-0003-2965-3104

REFERENCES

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martinez-Ortiz, I., & Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers in Education*, *141*, 103612. https://doi.org/10.1016/j.compedu.2019.10361

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

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–16. https://doi.org/http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief* (pp. 1–57. Retrieved from). Washington, DC: SRI International. https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, *28*(3), 771–780. https://doi.org/10.1016/j.chb.2011.11.020

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

Clark, D. B., Martinez-Garza, M. M., Biswas, G., Luecht, R. M., & Sengupta, P. (2012). Driving assessment of students' explanations in game dialog using computer-adaptive testing and hidden Markov modeling. In *Assessment in Game-Based Learning* (pp. 173–199). https://doi.org/10.1007/978-1-4614-3546-4_10

Comunidad de Madrid. (2016). Datos y cifras de la Educación 2016/2017.

Deepak, E., Pooja, G. S., Jyothi, R. N. S., Kumar, S. V. P., & Kishore, K. V. (2016). SVM kernel based predictive analytics on faculty performance evaluation. *2016 International Conference on Inventive Computation Technologies (ICICT)*, 1–4. https://doi.org/10.1109/INVENTIVE.2016.7830062

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*(x), 155–161. https://doi.org/10.1.1.10.4845

e-UCM. (2012). First-Aid Game. Retrieved from http://first-aid-game.e-ucm.es/

Frederick-Recascino, C., Liu, D., Doherty, S., Kring, J., & Liskey, D. (2013). Articulating an experimental model for the study of game-based learning. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): Vol. 8018 LNCS* (pp. 25–32). https://doi.org/10.1007/978-3-642-39226-9_4

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 *Learning, Design, and Technology* (pp. 1–29). https://doi.org/10.1007/978-3-319-17727-4_21-1

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2016). A practical guide to support vector classification. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Instituto Nacional de Evaluación Educativa—Ministerio de Educación, C. y D. (2017). Sistema Estatal de Indicadores de la Educación Edición 2016. Retrieved from http://www.mecd.gob.es/dctm/inee/indicadores/2016/17768–sistemaestatalindicadores2016-27-3-2017.pdf?documentId=0901e72b824643f0%5Cnhttp://www.mecd.gob.es/inee/sistema-indicadores/Edicion-2016.html

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, 31–40.

Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281. https://doi.org/10.1504/ijdats.2011.041335

Manero, B., Torrente, J., Freire, M., & Fernández-Manjón, B. (2016). An instrument to build a gamer clustering framework according to gaming preferences and habits. *Computers in Human Behavior*, *62*, 353–363. https://doi.org/10.1016/j.chb.2016.03.085

Marchiori, E. J., Ferrer, G., Fernandez-Manjon, B., Povar-Marco, J., Suberviola, J. F., & Gimenez-Valverde, A. (2012). Video-game instruction in basic life support maneuvers. *Emergencias*, *24*(6), 433–437.

Michael, D. R., & Chen, S. L. (2005). Serious games: Games that educate, train, and inform. *Education*, October, *31*, 1–95. https://doi.org/10.1145/2465085.2465091

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

Perez Colado, I., Perez Colado, V., Martínez-Ortiz, I., Freire, M., & Fernandez-Manjon, B. (2017). uAdventure: The eAdventure reboot—Combining the experience of commercial gaming tools and tailored educational tools. *IEEE Global Engineering Education Conference (EDUCON)*, 1754–1761. Retrieved from http://www.e-ucm.es/drafts/e-UCM_draft_304.pdf

Perez-Colado, I. J., Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Simva: Simplifying the scientific validation of serious games. *9th IEEE International Conference on Advanced Learning Technologies (ICALT)*.

Perez-Colado, I. J., Alonso-Fernández, C., Freire-Moran, M., Martinez-Ortiz, I., & Fernández-Manjón, B. (2018). Game learning analytics is not informagic! *IEEE Global Engineering Education Conference (EDUCON)*.

Petri, G., & Gresse von Wangenheim, C. (2016). How to evaluate educational games: A systematic literature review. *Journal of Universal Computer Science*, *22*(7), 992–1021. https://doi.org/10.3217/jucs-022-07-0992

Petri, G., & Gresse von Wangenheim, C. (2017). How games for computing education are evaluated? A systematic literature review. *Computers & Education*, *107*, 68–90. https://doi.org/10.1016/j.compedu.2017.01.004

Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458–472. https://doi.org/10.1016/j.compedu.2013.06.009

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

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

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Shute, V., & Kim, Y. J. (2014). Formative and stealth assessment. In *Handbook of research on educational communications and technology: Fourth edition* (pp. 311–321). https://doi.org/10.1007/978-1-4614-3185-5_3

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

Shute, V. J., & Moore, G. R. (2017). Consistency and validity in game-based stealth assessment. In *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective*.

Spalek, T. M., & Hammad, S. (2005). The left-to-right bias in inhibition of return is due to the direction of reading. *Psychological Science*, *16*(1), 15–18. https://doi.org/10.1111/j.0956-7976.2005.00774.x

How to cite this article: Alonso-Fernández C, Martínez-Ortiz I, Caballero R, Freire M, Fernández-Manjón B. Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *J Comput Assist Learn*. 2019;1–9. https://doi.org/10.1111/jcal.12405

APPENDIX A: | LIST OF GAME INTERACTION VARIABLES

Table A1 provides the full list of variables derived from the xAPI-SG statements collected from students' gameplays. These variables are used as input for the prediction models. Table A1 provides the variables names, types, and detailed description.

TABLE A1 Variables selected from game interaction xAPI statements

Variable name	Type	Description
gameCompleted	Binary (true, false)	True if learner completed the game; false otherwise
Score	Numerical in range [0, 10]	Total score obtained in the game
maxScoreCP	Numerical in range [0, 10]	Maximum score obtained in "chest pain" level
maxScoreU	Numerical in range [0, 10]	Maximum score obtained in "unconsciousness" level
maxScoreCH	Numerical in range [0, 10]	Maximum score obtained in "choking" level
firstScoreCP	Numerical in range [0, 10]	First score obtained in "chest pain" level
firstScoreU	Numerical in range [0, 10]	First score obtained in "unconsciousness" level
firstScoreCH	Numerical in range [0, 10]	First score obtained in "choking" level
timesCP	Integer	Number of times student completed "chest pain" level
timesU	Integer	Number of times student completed "unconsciousness" level
timesCH	Integer	Number of times student completed "choking" level
int_patient	Integer	Number of interactions with patient (game character, NPC)
int_phone	Integer	Number of interactions with phone (game element)
int_saed	Integer	Number of interactions with defibrillator (game element)
failedEmergency	Binary (true, false)	True if learner failed, at least once, the question about the emergency number; false otherwise
failedThrusts	Binary (true, false)	True if learner failed, at least once, the question about the number of abdominal thrusts per minute; false otherwise
failedHName	Binary (true, false)	True if learner failed, at least once, the question about the name of Heimlich manoeuvre; false otherwise
failedHPosition	Binary (true, false)	True if learner failed, at least once, the question about the initial position for Heimlich manoeuvre; false otherwise
failedHHands	Binary (true, false)	True if learner failed, at least once, the question about the hand position for Heimlich manoeuvre; false otherwise