

This manuscript is a post-print version of the document published in:

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. Improving the prediction of learning outcomes in educational platforms including higher level interaction indicators. Expert Systems (In press). 2018. <https://doi.org/10.1111/exsy.12298>

<https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.12298>

© 2018 Wiley Online Library

ARTICLE TYPE

Improving the Prediction of Learning Outcomes in Educational Platforms including Higher Level Interaction Indicators

José A. Ruipérez-Valiente¹ | Pedro J. Muñoz-Merino² | Carlos Delgado Kloos²

¹Massachusetts Institute of Technology (MIT),
77 Massachusetts Ave, Cambridge MA, USA

²Universidad Carlos III de Madrid, Avenida
Universidad 30 Leganes, Spain

Correspondence

*José A. Ruipérez-Valiente. Email:
jruipere@mit.edu

Present Address

600 Technology Square, 2nd floor, Office
2027B Cambridge (MA), United States, 02139

Summary

One of the most investigated questions in education is to know which factors or variables affect learning. The prediction of learning outcomes can be used to act on students in order to improve their learning process. Several studies have addressed the prediction of learning outcomes in Intelligent Tutoring Systems (ITSs) environments with intensive use of exercises, but few of them addressed this prediction in other web based environments with intensive use not only of exercises but also e.g. of videos. In addition, most works on prediction of learning outcomes are based on low level indicators like number of accesses or time spent in resources. In this paper we approach the prediction of learning gains in an educational experience using a local instance of Khan Academy platform with an intensive use of exercises and taking into account not only low level indicators but also higher level indicators such as students' behaviours. Our proposed regression model is able to predict 68% of the learning gains variability with the use of six variables related to the learning process. We discuss these results providing explanation of the influence of each variable in the model and comparing these results with other prediction models from other works.

KEYWORDS:

Decision support system, Educational data mining, Learning analytics, Predictive modelling

1 | INTRODUCTION

In the field of education, there is extensive work for trying to predict learning outcomes in educational experiences; these works are often classified inside the Educational Data Mining (EDM) field of research. EDM is considered one of the key research fields on education for the next years. A revision of the current and future state of EDM can be seen in the review performed by R. S. J. D. Baker and Yacef (2009). The raw data generated by different Virtual Learning Environments (VLEs) have been used for a wide variety of purposes, such as to predict course dropouts in MOOCs (Kloft, Stiehler, Zheng, & Pinkwart 2014) and also in the context of high school education (Márquez-Vera et al. 2015), as decision support for college admissions (Janecek & Haddawy 2007), to predict if students are going to surpass a course or not (Delgado Calvo-Flores, Gibaja Galindo, Pegalajar Jiménez, & Pérez Piñeiro 2006), to predict the major that a student is going to pick, before the student actually enrolls in college courses (Pedro & Ocumpaugh 2014), to adapt learning environments to students' cognitive styles (Guo & Zhang 2009), to provide information about the performance of groups in collaborative learning environments (Perera, Kay, Koprinska, Yacef, & Zaane 2009), for group formation depending on students' learning styles (Dwivedi & Bharadwaj 2015), to support the recommendation elicitation process in online learning environments (Santos & Boticario 2015) or to predict the score of a test before actually doing it (Feng, Heffernan, & Koedinger 2006; Z. A. Pardos, Gowda, Ryan, & Heffernan 2010). These results have a direct impact on creating tools that can help to improve these learning experiences. For example, the implementation of recommendations systems which can try to modify bad behaviours for learning and promote the good ones: e.g. if hint abusing behaviour is found

to be bad for learning, the system could send a warning advising the student not to abuse hints and make a better use of them. A complete description of a system which uses predictive analysis in order to improve the learning process can be found in the reference by Essa and Ayad (2012); their work describes Student Success System (S3) where they identify students at risk by applying prediction modeling, they design the interventions to mitigate that risk and finally they close the feedback loop by checking the effectiveness of the applied intervention.

The increased use of Massive Open Online Courses (MOOCs) encourages the use of predictive models to try to foretell the future and take early decisions. The massiveness of these courses makes it very difficult for teachers to keep track of students, thus instructors can use the help provided by decision making tools based on data. In addition, MOOCs usually have a high number of students, which makes them a perfect scenario to apply predictive analysis techniques. As an example, the work carried out by Brinton et al. (2014) analyses data from more than 100.000 distinct students from Coursera MOOCs, which is a data sample size hardly available in any other educational contexts. The specificities of MOOC platforms are different from other VLEs such as ITSs or Learning Management Systems (LMSs). Some aspects that are usually different are the format of the course, the intensive use of videos or the type of exercises. In addition, the data stream used in each prediction experiment can be different. Some examples have used variables including help-seeking actions of students (Feng, Beck, Heffernan, & Koedinger 2008; Feng et al. 2006) from a survey with the objective of predicting their performance (Bekele & Menzel 2005), reading patterns (Mills, Bosch, Graesser, & D'Mello 2014) or facial expression (Grafsgaard, Wiggins, & Boyer 2014; Jaques, Conati, Harley, & Azevedo 2014; Muldner, Burleson, & Vanlehn 2010). Therefore, the comparison between different works must take into account the selected variables, which are in some cases dependent on the platform.

This work aims at proposing a predictive model of learning gains based on variables in a MOOC educational environment, in which there is an intensive use of videos and exercises. In our experiences we use MOOC technologies but the access is restricted to a predefined number of students in what is so called Small Private Online Courses (SPOCs). Our motivation in this work is to be able to design a prediction model that can make a good estimation of how much students are going to learn in these types of educational experiences. We want to gain insight regarding some questions such as: Which are the variables that have the higher prediction power in this context? Are they related to videos or exercises? How much variability of the learning gains can be explained with the prediction model? Which behaviours are bad for learning achievement? Being able to provide an answer to these questions, would help the community deepen into which are the actions of students using VLEs that really lead into a bigger learning achievement.

2 | RELATED WORK

In this first stage of the related work we review some studies which apply prediction methods in learning environments, that do not try to predict learning achievement directly, but have other educational objectives. There is research to predict if a student is going to enroll in a certain type of major in college (Pedro & Ocumpaugh 2014) or to be able to predict if a student is going to quit reading while learning from instructional texts (Mills et al. 2014). Other works have approached the prediction of sentiments in education. For example, the study by Jaques et al. (2014) predicts emotions of students while using MetaTutor, using eye-tracking sensors as data inputs. Another research work tries to predict moments of delight dubbed as "yes;" with both interaction and physiological sensor features (Muldner et al. 2010). Moreover, Wang and Mitrovic (2002) approached the prediction of the number of errors in each submission of students using SQL-Tutor with high prediction accuracy. One last interesting work is the possibility of predicting the quality of the tasks developed in peer assessments through quantitative ratings and descriptive comments (Yu & Wu 2013). Although these works do not aim to predict test scores or learning gains, they help to understand that the prediction objectives in education can be very diverse also.

More closely related to our study we can find several works in the literature which deal with the prediction of a post-test transfer or learning gains. Both cases are quite similar as learning gains depend on the post-test score. In this direction, there are several research works on the ASSISTment system (Anozie & Junker 2006; Feng et al. 2008 2006; Kelly, Arroyo, & Heffernan 2013) which objectives are similar to ours; that is the prediction of some student performance indicator at the end of a course, by using the data generated by the system. The results of these works developed on ASSISTment utilize variables related to help-seeking behaviour and others more general about time or percentage of correct items. In addition to this type of variables, in our study we have made use of more complex variables such as for students' behaviour. Others similar works are based on an ITS for College Genetics where they also try to predict learning outcomes (R. S. Baker, Gowda, & Corbett 2011; R. S. J. D. Baker et al. 2010; Corbett, Kauffman, Maclaren, Wagner, & Jones 2010). There has been research on this College Genetics ITS which reported that the developed detector needed only a limited amount of data (around the first 20% of a students' data) in order to predict with reasonably accuracy (R. S. J. D. Baker et al. 2010); this is very interesting as it would allow to intervene in the early stages of a course. Another work on the College Genetics ITS compares several Bayesian Knowledge Tracing variants in order to see which one of them predicts better post-test performance (R. S. Baker et al. 2011). These research works also considered some variables that we have included in our research such as the total amount of time or average number of attempts. However, as ITS environments have different specificities as MOOC platforms, then the same variables cannot be applied in both environments and the effect as predictors of variables might be different because of the change of educational environment.

Another important question that has been addressed in the field of prediction modeling on education is about which techniques or algorithms should be used. We can even find several papers which focus is to compare different techniques or variants of the same algorithm with the purpose of finding which one is the more effective to predict learning outcomes (R. S. Baker et al. 2011; Janecek & Haddawy 2007; Koutina & Kermanidis 2011). The research by Kotsiantis (2012) performs a review of the different machine learning techniques for educational purposes. We have used linear regression as other works do as well (Feng et al. 2008 2006; Grafsgaard et al. 2014; Kelly et al. 2013), because we expect a linear relationship between the selected variables and students' learning gains. Other authors used different methods such as Bayesian Knowledge Tracing model (R. S. Baker et al. 2011), 1-NN (Koutina & Kermanidis 2011), Neuronal Networks such as Radial Basis Functions (Delgado Calvo-Flores et al. 2006) or machine learning techniques (Mills et al. 2014). Another interesting approach is to ensemble different prediction methods to achieve more robust results (Z. A. Pardos et al. 2010).

Our approach selected different types of variables such as for progress in exercises and videos, time invested or students' behaviours while solving exercises; these variables can be retrieved after processing the raw data generated by the students while interacting with the learning environment. Other studies used different types of variables such as Z. Pardos and Baker (2014), where the predictor variables represent affective states in order to predict the test scores. Other types of variables that can be used are gestures and postures, which are helpful to detect sentiments or even improve learning gains predictions (Grafsgaard et al. 2014). Finally, researchers also need to develop tools where these predictions can be useful to act on the learning process, such as recommendation or warning systems (Hu, Lo, & Shih 2014).

Most of the works presented in this section are based on ITSs, which usually have some differences with MOOC platforms, such as the intensive use of videos. MOOCs are recent and there are not as many works on prediction as on the ITS field. One of the most problematic issues of MOOCs is the high dropout ratio. We can find prediction studies about the dropout ratio using data from edX (Balakrishnan & Coetzee 2013) and also Coursera MOOCs (Kloft et al. 2014; Rosé & Siemens 2014); these studies try to reveal which are the variables that can predict quitting, so that instructors can intervene before that happens. Another key feature of MOOCs is the social activity and the prediction of how social activity evolves and which variables are important (Brinton et al. 2014). An interesting approach is to predict if students are going to solve correctly a question using video-watching stream data (Brinton & Chiang 2015). Although transfer tests and prediction models for learning gains have been addressed in different works within ITS context, to the best of our knowledge this approach has not been carried out in MOOC platforms yet. In previous work we have already proposed some of the indicators that we use in this work explaining the specificities of more complex behaviors such as video avoidance or unreflective user (Muñoz-Merino, Ruipérez Valiente, & Kloos 2013). We have also performed extensive relationship mining between the different variables in previous studies. However, our approach in this work is the analysis of how these variables are related to learning gains and how to use them to predict the learning achievement of students.

3 | METHODOLOGY

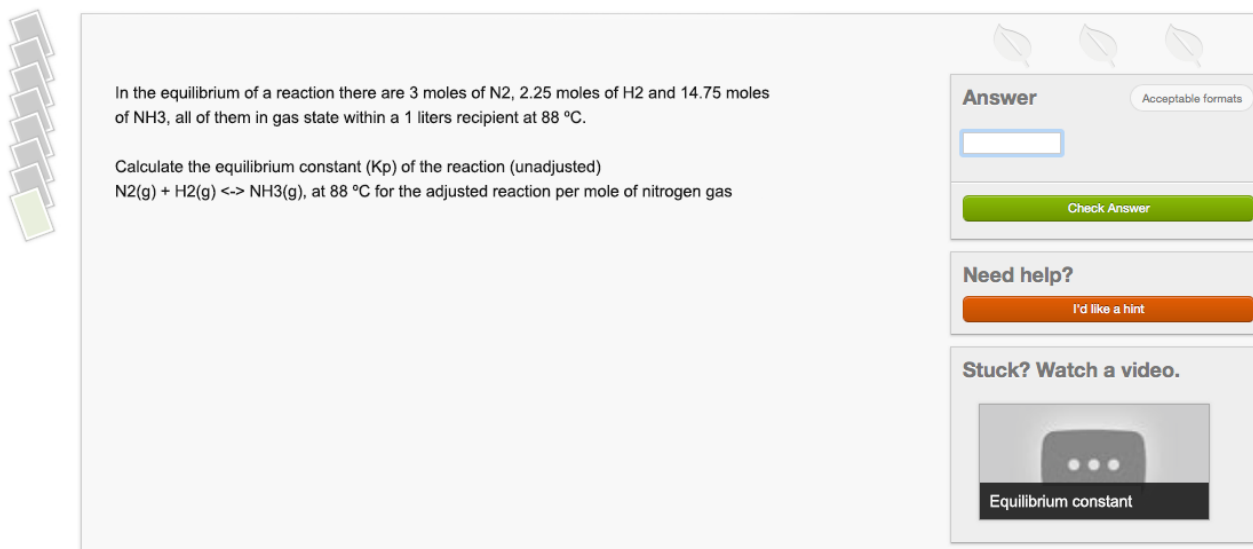
3.1 | Learning context

The learning context of this experience is located within the 0-courses for freshmen students, which take place at Universidad Carlos III de Madrid. These courses are for students starting an engineering degree at the university, so that they can review the concepts that are required before the beginning of the academic year. In the August month of the years 2012 to 2014 the university has provided MOOC technology to support these courses; in these last courses UC3M has enabled a Khan Academy instance where students could access and interact with the videos and exercises of the course. The idea is that students can review the concepts at home using Khan Academy, before actually attending the 0-courses face-to-face lessons, which take place in the first week of September.

In this experience, the main educational resources were exercises and videos. An example of a chemistry exercise from our instance is shown in figure 1. Students can ask for hints, attempt to answer the question several times until solving it correctly and watch the video related to that exercise if they are struggling. An important detail is that students need to solve several exercises correctly of the same type in order to acquire a proficiency level in that skill. Most of the exercises are parametric, which means that when a student correctly solves an exercise, the variables of the exercise will change but the statement will remain the same. In addition, Khan Academy incorporates other features such as gamification elements or learning analytics visualizations.

This research has been conducted in the chemistry and physics courses of 2014 as these courses are the only ones where we have a pre-test and post-test to see a transfer of learning. Courses were composed of 51 exercises and 24 videos for chemistry, and 33 for both exercises and videos for physics. The total amount of students who logged into the Khan Academy platform at least once was 156 for physics and 69 for chemistry (the total number of students enrolled to these courses is higher, but not all of them accessed the platform).

Practicing Equilibrium constant | in Equilibrium constant



In the equilibrium of a reaction there are 3 moles of N₂, 2.25 moles of H₂ and 14.75 moles of NH₃, all of them in gas state within a 1 liters recipient at 88 °C.

Calculate the equilibrium constant (K_p) of the reaction (unadjusted)
 N₂(g) + H₂(g) <=> NH₃(g), at 88 °C for the adjusted reaction per mole of nitrogen gas

Answer Acceptable formats

Check Answer

Need help?

I'd like a hint

Stuck? Watch a video.

Equilibrium constant

FIGURE 1 Exercise example of our chemistry Khan Academy instance.

3.2 | ALAS-KA

ALAS-KA (Add-on of the Learning Analytics Support for the Khan Academy) is a plug-in that enables a new visualization dashboard within the Khan Academy environment. This tool enables a set of new indicators that are not available in the default interface of Khan Academy, for example related to cognitive behavior, the time distribution by students, efficiency and effectiveness with different educational resources and more. Each of these indicators contains individual visualizations per each student separately compared to average values, as well as global aggregations for the entire class. Instructors have full access to all information for personal tracking of each student separately, but also to global aggregations of the class for more general trends. Students were allowed to access all their own information and visualizations for self-awareness. For a detailed functionality and indicators, it is possible to consult a previous publication about ALAS-KA (Ruipérez-Valiente, Muñoz-Merino, Leony, & Delgado Kloos 2015). Additionally, a video with a usage example of the tool is available online¹.

During the educational context described in Subsection 3.1, ALAS-KA was enabled, which permitted the instructors of the course to do a closer monitorization of their students and to perform a just-in-time detection of problematic contents of the course, as well as help with keeping the engagement of students high by providing additional information regarding their progress. Most of the indicators that we describe in Subsection 3.4 and use for the model in Section 4 have been computed by ALAS-KA.

3.3 | Dataset and experiment

As we want to measure the learning outcome we use learning gains (LGs) of students during their interaction with the platform, we implemented a pre-test (before interacting with the platform) and a post-test (after finishing the interaction with the platform, i.e. before the face to face sessions). The tests from physics course had 10 questions each whilst the tests in chemistry had 21 questions each, as the contents of the chemistry course were broader, more questions were necessary to cover all the concepts. The difficulty of both tests was similar as the problems were randomly pulled from a pool of questions of equal hardness.

We define a student learning gain by obtaining the difference between post-test minus pre-test (LG = post - pre). Both tests are scored from 0 to 100 points; therefore learning gains can range from -100 to 100. The contents of the course (exercises and videos) were disabled until they finished the pre-test, so that we can really measure students' initial knowledge before using the platform. The post-test was done at the end of August before the 1st of September when face-to-face lessons started. In addition, when processing the considered variables, we only take into account the interaction of students until the post-test was done, the rest was discarded.

¹<https://www.youtube.com/watch?v=vDs1tt7siBA>

A total of 163 students in physics and 77 in chemistry attempted the pre-test, but just 48 students in physics and 30 in chemistry finished both tests. In addition not all the students who finished both tests have been included in the analysis. We incorporated a condition where students needed to spend at least 30 seconds multiplied by the number of questions interacting with the test. The objective is to remove those students who were randomly answering the test and did not make enough reflection on the different questions. Thus a time threshold of 5 minutes for physics and 10 minutes and a half for chemistry was applied to incorporate the student to the data sample. Under these conditions, the total number of students that were introduced to the prediction model are 25 for chemistry and 44 for physics, which makes a total amount of 69 students.

3.4 | Considered variables

In previous work we proposed a prediction model which took into account low level indicators such as *avg_hints* or *avg_attempts* count required by students to solve an exercise (Ruipérez-Valiente, Muñoz-Merino, & Delgado Kloos 2015). In that previous work the variables that we took into account for the prediction model were the next.

- *pre_test_score*: the score obtained by the student in the pre-test (from 0 to 100).
- *pre_test_time*: time required by the student to complete and submit the pre-test (minutes).
- *correct_exercises*: percentage of correct exercises of the total number the student attempted to solve (from 0 to 100).
- *exercises_solved_once*: percentage of different types of exercises that were solved correctly at least once (from 0 to 100).
- *proficient_exercises*: percentage of exercises in which the student has acquired a proficiency level (from 0 to 100).
- *avg_hints*: average number of hints asked by the student.
- *avg_attempts*: average number of attempts that a student makes trying to solve an exercise.
- *avg_video_progress*: average progress by the student in each video (from 0 to 100).
- *videos_completed*: percentage of videos completed by the student (from 0 to 100).
- *total_time*: total time spent by the student in exercises and videos (minutes).
- *exercise_time*: time spent by the student solving exercises (minutes).
- *video_time*: time spent by the student watching videos (minutes).

These variables are closely related to some of other works such as (Feng et al. 2008 2006). We obtained a regression model that could predict 57.4% of the learning gains variability by using the *pre_test_score*, *avg_attempts*, *total_time* and *proficient_exercises* variables (Ruipérez-Valiente, Muñoz-Merino, & Delgado Kloos 2015). In this work we have tried to improve our prediction model by using new learning variables that could provide of a better prediction power; but we have also taken into account the previous set of variables. Some of these variables try to improve the ones that were used previously and others are completely new and related to behaviors of students while interacting with the platform which represent complex indicators. We present next the new set of variables that have been taken into consideration for the new prediction model:

- *optional_activities*: this variable measures the number of optional activities (such as setting up an avatar or learning goals) that have been used by the student (Ruipérez-Valiente, Muñoz-Merino, Delgado Kloos, Niemann, & Scheffel 2014).
- *correct_exercises_no_help*: percentage of exercises correctly solved by the student of the total number that the student attempted to solve, without the use of hints and in their first attempt (it is possible to attempt several times if the student fails to answer correctly).
- *exercise_effectiveness*: this is a specific variable with uses a non-linear function to measure the total progress of students in exercises, taking into account that most of the exercises were parametric (Muñoz-Merino, Ruipérez-Valiente, Alario-Hoyos, Pérez-Sanagustín, & Delgado Kloos 2015).
- *video_effectiveness*: this is a specific variable which also uses a non-linear function to measure the total progress of students in videos, taking into account the specificities of the videos developed for these courses (Muñoz-Merino et al. 2015).
- *mean_daytime*: this measure represents the mean of the time spent each day. It takes into account the data from the pre-test date to the post-test date.

- *variance_daytime*: this measure represents the variance of the time spent each day. It takes into account the data from the pre-test date to the post-test date.
- *efficiency_time*: this variable provides an efficiency measure which takes into account the amount of time that students needed, in order to solve their exercises correctly.
- *follow_recommendations*: this measure provides the percentage of exercises that were accessed by the student via a Khan Academy's resource recommendation.
- *forgetful_user*: this variable provides information about the percentage of exercises that students failed to solve after solving an exercise of the same type correctly.
- *exercise_abandonment*: percentage of exercises that were started but the student never achieved proficiency in them.
- *video_abandonment*: percentage of videos that were started by the student but never completed.
- *hint_avoidance*: indicator about users who failed to correctly solve exercises and still they do not ask for hints.
- *video_avoidance*: indicator about users who failed to correctly solve exercises and still they do not watch the video which is associated to that exercise.
- *hint_abuse*: indicator about students who ask for too many hints without reflecting on the exercise statement or previous hints.
- *unreflective_user*: indicator about students who attempt to solve an exercise too many times without reflecting.

We believe that with the inclusion of the new variables we will be able to improve the previous prediction model. Some of the variables try to improve the ones that were considered in previous work and others introduce new learning behaviors that were not considered in previous work. For more information about the variables that we have taken into account for this research study it is possible to consult previous works (Muñoz-Merino et al. 2013; Ruipérez-Valiente, Muñoz-Merino, Leony, & Delgado Kloos 2015). In addition, other works have addressed some of the help-seeking behaviors that we are also taking into consideration, such as Aleven, McLaren, Roll, and Koedinger (2004); they performed a correlation analysis of some variables that are similar to ours such as help abuse or help avoidance with learning gains. Since we are trying to predict learning gains using mainly variables obtained from the interaction of students with the platform, demographics variables are not included in this study as a difference with other studies that do use variables such as gender or nationality.

4 | PREDICTION MODEL AND DISCUSSION

The first action undertaken in order to design this model was to review the state of the art to check on works which used similar variables, in order to have an idea of which variables might have a potential impact on learning achievement. Next, we also performed an exploratory analysis by applying stepwise regression methods, in order to determine which variables had the highest impact on the prediction model. In addition, we also performed a correlation analysis between learning gains and all the considered variables. The objective of this previous analysis is to establish which are the most powerful variables that can be used to predict learning gains.

We selected a hierarchical method with three entry steps and a total of six independent variables (introducing two of them in each step). We selected this method as we ran an exploratory analysis (as explained in the previous paragraph) before designing the model, and we used that outcome to define the model. Our rationale to follow this methodology is avoiding over-fitting that can come from running algorithms that perform automatic selection of features, specially since our data sample is not big. In the first step, we introduced the two variables which are the same that we used in our previous research, as these variables are still significant predictors. This decision was also supported the work of Feng et al. (2006), where they made use of these variables too in their research. In the second step we improved the variables that we used in our previous work. In our previous work we used *proficient_exercises* and *total_time* (which are related to progress in exercises and time invested in exercises and videos by the student) in the second step. In this research we have improved these variables and substitute them by *correct_exercises_no_help* and *mean_daytime*, which provide of a more powerful prediction power despite the information that they are transmitting is similar. In the last step we introduced the last two variables which are related to the behaviour of students while interacting with the platform. As we have a limited amount of cases we should not introduce all the predictors that we wanted, in order to comply with dummy rules such as 10 cases per predictor variable. Therefore we decided to add two new variables related to the behaviours of students; one collects the variables which have a significant prediction power measuring increment of the learning gains prediction and the other one has a decremental influence. The new variables introduced in the third step are the following:

- *exercise_video_abandonment*: this variable combines both *exercise_abandonment* and *video_abandonment*, as they have a significant influence on the prediction of the learning gain. These variables have an incremental influence on the prediction of the learning gain; this is interesting as we could guess that students who abandon exercises and videos would probably learn less.
- *negative_behaviors*: this variable combines *follow_recommendations*, *forgetful_user* and *unreflective_user*. An interesting detail is that other behavioural variables such as *video_avoidance*, *hint_avoidance* or *hint_abuse* were not as significant as the others, thus they were left outside the model. These variables have a decremental influence on the prediction model; this makes sense in the case of for *forgetful_user* and *unreflective_user* but no so much about *follow_recommendations*.

We discuss more about the effect of each variable in the model later. Table 1 shows the summary of the three models. The first model with the same two variables that we used in our previous research provides a R^2 of 0.481. The second models add two new variables rising up to 0.616; the use of the new variables in the second models involves a relative improvement of the R^2 of 0.042 points with respect the model presented in our previous work. Finally, the third model includes the two variables related to students' behaviour and provides a R^2 of 0.68, which means that our final model is able to predict a 68% of the learning gain's variability. This last result provides of a relative improvement of the R^2 of 0.106 with respect our previous work thanks to the improvement of the variables and the addition of new ones related to students' behaviour. The standard error of prediction is 13.3; this means that when making a prediction the average deviation from the real value is of 13.3 points. A first impression about the importance of each one of them can be obtained at table 2 by checking the standardized coefficients. Equation 1 shows the complete prediction formula; please note that this formula is represented with un-standardize coefficients.

TABLE 1 Model summary of the linear regression model.

Model	R	R Square	Std. Error of the Prediction
1	0.693	0.481	16.42
2	0.785	0.616	14.34
3	0.825	0.68	13.3

TABLE 2 Unstandardized and standardized coefficients of the regression models.

Model	Independent Variable	Un-std. Coeff.		Std. Coeff.
		B	Std. Error	Beta
1	Constant	38.556	7.88	
	<i>pre_test_score</i>	-0.601	0.84	-0.655
	<i>avg_attempts</i>	4.093	3.149	0.119
2	Constant	14.485	8.991	
	<i>pre_test_score</i>	-0.646	0.076	-0.703
	<i>avg_attempts</i>	5.362	2.776	0.156
	<i>correct_exercises_no_help</i>	0.271	0.106	0.224
	<i>avg_day_time</i>	0.557	0.200	0.231
3	Constant	13.615	9.734	
	<i>pre_test_score</i>	-0.668	0.071	-0.727
	<i>avg_attempts</i>	6.426	3.142	0.187
	<i>correct_exercises_no_help</i>	0.392	0.104	0.324
	<i>avg_day_time</i>	0.824	0.230	0.342
	<i>exercise_video_abandonment</i>	0.143	0.097	0.155
	<i>negative_behaviors</i>	-0.721	0.223	-0.264

$$\begin{aligned}
 LG = \{ & 13.615 - 0.668 * pre_test_score + 6.426 * avg_attempts \\
 & + 0.392 * correct_exercises_no_help + 0.824 * avg_day_time \\
 & + 0.143 * exercise_video_abandonment - 0.721 * negative_behaviors \}
 \end{aligned}
 \tag{1}$$

Next we analyse each one of the model predictors separately:

- *pre_test_score*: This variable represents the most powerful predictor. The explanation of the negative sign is related to the fact that if the initial knowledge of students is very high, it will be harder to improve that knowledge. For example, it is hard that a student who scores 90 in the pre-test achieves a 100 score in the post-test. However, it would be very probable that a student who has scored 0 at the pre-test, will score 10 or higher at the post-test after using the platform. The higher value of the pre-test, the harder is to increase the post-test score with respect to the pre-test. For every point in the pre-test, the predicted learning gain decreases 0.668 points.
- *avg_attempts*: The average number of attempts that students made trying to solve an exercise reports a positive effect towards predicting a learning gain. The higher the average number of attempts the better. A possible hypothesis would be that students who cannot solve exercises, do not even attempt to solve them and just leave, so they do not increase learning. In addition, students who make a lot of attempts might learn by error and repetition and thus they can obtain a higher learning gain in this process than just students that answer the question directly. For every unit that the average number of attempts increases, the predicted learning gain will raise 6.426 points.
- *correct_exercises_no_help*: The percentage of correct exercises without use of hints and answering correctly at the first attempt represents one of the most important predictors of the model. This makes sense as the more exercises students are able to solve without help, more likely is that their knowledge is higher. For every point that the variable increases, the predicted learning gain will increase 0.392 points.
- *avg_day_time*: The average number of minutes spent by the student each day is the second most important predictor of the model. It makes sense that the bigger is the amount of time invested by students in the platform, the higher is going to be the increment of their knowledge. However there might be cases in which this relationship does not apply. For every minute that the average time per day increases, the predicted learning gain will increase 0.824.
- *exercise_video_abandonment*: This variable has the lowest weight on the prediction model; however it also helped to improve the prediction power as it probably provides of variability that was not provided by other variables. Surprisingly, if the amount of exercises and videos that students abandon increase, the predicted learning gain will also increase. A possible explanation that students that have high abandonment ratios might be abandoning those resources because they already have that knowledge, thus they will score high on the post-test later and that will result in a learning gain increment. For every point that this variable increases, the predicted learning gain will increase 0.143.
- *negative_behaviors*: The highest is this variable the lowest is going to be the predicted learning gain. This relationship makes sense for *forgetful_user* and *unreflective_user*, as we would think that these behaviors do not represent good actions for learning. Students who forget how to solve exercises mean that they are not really correctly acquiring the knowledge and unreflective students do not have the knowledge to solve the exercises and are not reflecting on their mistakes. However the relationship with *follow_recommendations* is not quite straightforward. A possible hypothesis could be that students who follow recommendations, do not have a good background knowledge about the topics covered and they are going step by step; whereas students with a good background will jump from one topic to another, going to those topics which are more appealing for them.

We have used only 6 predictor variables although we considered much more in our analyses because the number of cases that we have in our data sample is 69. One important detail is the importance of exercises and videos to the prediction of learning gains as these experiences had an intensive use of both of them.

Taking into account all the variables that have been incorporated into the model, and also taking into consideration that *exercise_video_abandonment* as well as *negative_behaviors* include 2 and 3 different variables respectively, a total number of 6 variables are related to exercise activity (*avg_attempts*, *correct_exercises_no_help*, *avg_day_time*, *exercise_abandonment*, *forgetful_user*, *unreflective_user* and *follow_recommendations*) and 2 variables are related to videos (*avg_day_time* and *video_abandonment*). About these results we should take into account that there were more considered variables related to exercises, for example there were 8 behavioural variables, 6 of them related to exercises and only 2 to videos. Additionally we should add that despite variables related to progress in videos such as *videos_completed* or *video_effectiveness* were good learning gain predictors, however other variables related to progress in exercises such as *correct_exercises_no_help* or *proficient_exercises*, were more powerful predictors. In this case, a combination of variables related to exercise and video progress in a different variable (similar to what we did with the behavioural variables), did not result into a better prediction than the separate use of the variables.

All the assumptions of the regression model are fulfilled. There is no perfect multicollinearity between the predictor variables as our VIF values are below 10 and tolerance statistics are above 0.2; the independence of errors has been tested by the Durbin-Watson which is 2.123 (it needs to be close to 2). Additionally, the linearity and homoscedasticity assumptions are confirmed to be fulfilled by plotting the standardized residuals versus standardized predicted values. The normality of the residuals is tested by checking the histogram and normal probability plot of the residuals. We should also point out that there are zero cases with a standardized residual above ± 2 , which means that the model is well fitted and there are no outliers.

Therefore under these circumstances we should be able to say that the model would generalize well to predict other samples of the same population. We cannot however test this model with a cross-validation as the number of cases in the data sample is too small.

We can establish a comparison of results with some of the related works. Despite our research had several similarities with Feng et al. (2006), some variables are not the same because of the different nature of the learning environment, but others are the same such as *pre_test_score* or *avg_attempts*; additionally, we have considered behavioural variables which were not present at all in Feng's study.

The work by Kelly et al. (2013), which also made use of a linear regression analysis to predict standardized test scores, obtained a R^2 of 0.57; they used different variables except the average number of attempts. The work by Grafsgaard et al. (2014) makes use of posture and gesture data provided by sensors, they obtained a R^2 of 0.38 predicting learning gains; their work environment was a Java Tutor, therefore the variables of their model were different as it was a programming environment. The work by Anozie and Junker (2006) is reported a regression model which is able to account for the 63.7% of the variability; they also make use the *pre_test_score* and other variables related to time and percentage of correct exercises. One of the main differences of this work with others is that our learning environment was based in MOOC technologies and e.g. the course had an intense video activity, whereas none of the other works compared here used videos as part of their learning experience. Therefore, the considered variables change, as the context is different.

Furthermore we raise some questions that we would like to answer. First we would like to answer, how good would get the model if we could use more predictors? We have put a limit to the number of predictors since our data sample is small. A Backward Stepwise regression analysis reported that we could achieve a R^2 of 0.75 with the use of 16 of the considered variables; therefore we would have an upper limit of measuring 75% of the learning gains variability with the use of more variables. That would provide an improvement of 0.07 points with respect to our design using only 6 predictors (so adding many more variables would not represent a big improvement). Additionally, another interesting question is the effect of the pre-test variable on the prediction model. If we would remove this variable from the considered ones and repeat the Backward Stepwise regression, we could achieve only a R^2 of 0.481 with the use of 16 variables again; thus we can notice that the effect of using prior knowledge by using the pre-test score is highly important, as the influence of this variable in the prediction model cannot be covered by any of the rest of considered variables. A similar analysis of the influence of *pre_test_score* in their regression model was performed in Feng et al. (2006) with similar conclusions.

5 | CONCLUSIONS

In this work, we have approached the problem of predicting the increment of knowledge of students during their interaction with MOOC technologies, within a educational experience with intensive use of videos and exercises. We believe the results are good, as the model can predict 68% of the learning gains variability. In addition, the standard error of prediction considerably improves compared to the standard error of using the baseline prediction. Furthermore the model fulfilled all the assumptions for the regression analysis, thus it should generalize well under experiences with similar conditions. Our proposed predictive model uses six independent variables: one is the pre-test score and the other five have been retrieved from the low level data stored due to the interaction of students. This model represents advancement with respect our previous work as some of the variables have been improved and we have also taken into account behavioural variables. Most of the other similar works use more straightforward variables related only to time, correct exercises or number of accesses; so the introduction of these new behavioural variables can be useful for new researches.

For this model we have selected a multivariate linear regression to predict the learning gain. The main rationale for this decision is that as part of the research questions we want to analyze which variables have a positive or negative impact on the predicted learning gain. Therefore, if we were to use other well-known nonlinear models for this prediction, such as gradient boosted machines, k-nearest neighbors or random forests, these conclusions would not be possible, as there is not a linear relationship between the estimated model parameters and the predicted variable. As a handicap, a simple multivariate linear regression will not be the best performer in terms of quality metrics, such as the mean squared error. In this case, since this model has not been set up in a production environment, this is a drawback that we are able to accept. Therefore, these results can help in identifying which variables are related to learning achievement so that proper actions can be performed during the learning experience. For example, we learned that being unreflective was found to be bad for learning achievement, hence we could recommend students to reflect more on their previous attempts whenever this behaviour is detected.

One of the main limitations of this research is the important influence of the *pre_test_score* variable. We showed the high influence of this variable when removing it, and in experiences where it is not the case that the prior knowledge of students is available, the predictive model will perform worse. In this direction, a challenge would be to be able to acquire prediction results as high as the one hereby presented, but without using the prior knowledge of students, just with the use of variables obtained from the interaction of students with the platform. Furthermore, it also needs to be explored the outcomes of applying similar prediction analysis into other MOOC platforms such as Open edX or Coursera, and also with massive use of students.

As future work we would like to extend this model with even more behavioral parameters, such as the interest and behavior of students with gamification elements (Ruipérez-Valiente, Muñoz-Merino, & Delgado Kloos 2017). Currently, we are using only variables related to their interaction with the contents in the platform, but it would also be promising to follow a more mixed methods approach, for example by combining other information such as demographics and survey data into the model. It would also be feasible to consider multimodal approaches for this model, by introducing biometrics signals (heart rate, eye-gaze or voice). Finally, other issues arise as future steps: Is it possible to extrapolate most of these variables to different MOOC platforms? Can similar predictive analysis be performed with similar results on different MOOC platforms? We would like to be able to address these questions in future research.

ACKNOWLEDGEMENTS

This work has been supported by the “eMadrid” project (Regional Government of Madrid) under grant S2013/ICE-2715, the “RESET” project (Ministry of Economy and Competiveness) under grant RESET TIN2014-53199-C3-1-R and the European Erasmus+ SHEILA project under grant 562080-EPP-1-2015-BE-EPPKA3-PI-FORWARD.

References

- Aleven, V., McLaren, B. M., Roll, O., & Koedinger, K. (2004). Toward Tutoring Help Seeking; Applying Cognitive Modeling to Meta-Cognitive Skills. In *Seventh international conference on intelligent tutoring systems (its-2004)* (pp. 227–239). Springer Verlag. doi: 10.1007/b100137
- Anozie, N., & Junker, B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In *American association for artificial intelligence workshop on educational data mining (aaai-06)* (pp. 1–6). Boston, EEUU. doi: WS-06-05/WS06-05-001
- Baker, R. S., Gowda, S., & Corbett, A. (2011). Towards predicting future transfer of learning. In *Proceedings of 15th international conference on artificial intelligence in education* (pp. 22–30). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-21869-9_6
- Baker, R. S. J. D., Corbett, A. T., Gowda, S. M., Wagner, A. Z., MacLaren, B. a., Kauffman, L. R., ... Giguere, S. (2010). Contextual slip and prediction of student performance after use of an intelligent tutor. *Lecture Notes in Computer Science. User Modeling, Adaptation, and Personalization*, 6075, 52–63. doi: 10.1007/978-3-642-13470-8_7
- Baker, R. S. J. D., & 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. doi: <http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314>
- Balakrishnan, G., & Coetzee, D. (2013). *Predicting student retention in massive open online courses using hidden markov models* (Tech. Rep.). Electrical Engineering and Computer Sciences University of California at Berkeley.
- Bekele, R., & Menzel, W. (2005). A Bayesian Approach To Predict Performance Of A Student (BAPPS): A Case with Ethiopian Students. In *Artificial intelligence and applications* (pp. 189–194).
- Brinton, C., & Chiang, M. (2015). MOOC Performance Prediction via Clickstream Data and Social Learning Networks. In *Ieee conference on computer communications* (pp. 2299 – 2307). IEEE. doi: 10.1109/INFOCOM.2015.7218617
- Brinton, C., Chiang, M., Jain, S., Lam, H., Liu, Z., & Wong, F. (2014). Learning about social learning in MOOCs: From statistical analysis to generative model. *IEEE Transactions on Learning Technologies*, 7(4), 346 – 359. doi: 10.1109/TLT.2014.2337900
- Corbett, A., Kauffman, L., Maclaren, B., Wagner, A., & Jones, E. (2010). A Cognitive Tutor for Genetics Problem Solving: Learning Gains and Student Modeling. *Journal of Educational Computing Research*, 42(2), 219–239. doi: 10.2190/EC.42.2.e
- Delgado Calvo-Flores, M., Gibaja Galindo, E., Pegalajar Jiménez, M. C., & Pérez Piñeiro, O. (2006). Predicting students' marks from Moodle logs using neural network models. In *Current developments in technology-assited education* (pp. 586–590).
- Dwivedi, P., & Bharadwaj, K. K. (2015). E-Learning recommender system for a group of learners based on the unified learner profile approach. *Expert Systems*, 32(2), 264–276. doi: 10.1111/exsy.12061
- Essa, A., & Ayad, H. (2012). Improving student success using predictive models and data visualisations. In *Research in learning technology* (pp. 58–70). doi: <http://dx.doi.org/10.3402/rlt.v20i0.19191>

- Feng, M., Beck, J., Heffernan, N., & Koedinger, K. (2008). Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test?. In Baker & Beck (Eds.), *Proceedings of the 1st international conference on educational data mining* (pp. 107–116). Montreal.
- Feng, M., Heffernan, N., & Koedinger, K. (2006). Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In *Proceedings of 8th international conference, intelligent tutoring systems jhongli, taiwan* (pp. 31–40). Berlin: Springer-Verlag.
- Grafsgaard, J., Wiggins, J., & Boyer, K. (2014). Predicting Learning and Affect from Multimodal Data Streams in Task-Oriented Tutorial Dialogue. In J. Stamper, Z. Pardos, M. Mavrikis, & B. McLaren (Eds.), *Proceedings of the 7th international conference on educational data mining* (pp. 122–129).
- Guo, Q., & Zhang, M. (2009). Implement web learning environment based on data mining. *Knowledge-Based Systems*, 22(6), 439–442. doi: 10.1016/j.knosys.2009.06.001
- Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014, jul). Developing early warning systems to predict students' online learning performance. *Journal of Computers in Human Behavior*, 36, 469–478. doi: 10.1016/j.chb.2014.04.002
- Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *2007 37th annual frontiers in education conference - global engineering: Knowledge without borders, opportunities without passports* (pp. T2G-7–T2G-12). doi: 10.1109/FIE.2007.4417993
- Jaques, N., Conati, C., Harley, J., & Azevedo, R. (2014). Predicting Affect from Gaze Data during Interaction with an Intelligent Tutoring System. In *Proceedings 12th international conference, intelligent tutoring systems 2014, honolulu, hi, usa* (pp. 29–38). Springer International Publishing. doi: 10.1007/978-3-319-07221-0_4
- Kelly, K., Arroyo, I., & Heffernan, N. (2013). Using ITS Generated Data to Predict Standardized Test Scores. In *Educationaldatamining.org* (pp. 3–4). Memphis, Tennessee, USA.
- Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the emnlp 2014 workshop on analysis of large scale social interaction in moocs* (pp. 60–65).
- Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4), 331–344. doi: 10.1007/s10462-011-9234-x
- Koutina, M., & Kermanidis, K. L. (2011). Predicting Postgraduate Students' Performance Using Machine Learning Techniques. In *Artificial intelligence applications and innovations* (pp. 159–168). doi: 10.1007/978-3-642-23960-1_20
- Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2015). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, 33(1), 107–124. doi: 10.1111/exsy.12135
- Mills, C., Bosch, N., Graesser, A., & D'Mello, S. (2014). To Quit or Not to Quit: Predicting Future Behavioral Disengagement from Reading Patterns. In *Proceedings of 12th international conference, intelligent tutoring systems honolulu, hi, usa, 2014* (pp. 19–28). doi: 10.1007/978-3-319-07221-0_3
- Muldner, K., Bursleson, W., & Vanlehn, K. (2010). "Yes !": Using tutor and sensor data to predict moments of delight during instructional activities. In *Proceedings of 18th international conference, user modeling, adaptation and personalization. big island, hi, usa*. (Vol. 2010, pp. 159–170). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-13470-8_16
- Muñoz-Merino, P. J., RUIPÉREZ VALIENTE, J. A., & Kloos, C. D. (2013). Inferring higher level learning information from low level data for the Khan Academy platform. In *Proceedings of the third international conference on learning analytics and knowledge - lak '13* (pp. 112–116). New York, New York, USA: ACM Press. doi: 10.1145/2460296.2460318
- Muñoz-Merino, P. J., RUIPÉREZ-VALIENTE, J. A., Alario-Hoyos, C., Pérez-Sanagustín, M., & Delgado Kloos, C. (2015). Precise Effectiveness Strategy for analyzing the effectiveness of students with educational resources and activities in MOOCs. *Computers in Human Behavior*, 47, 108–118. doi: 10.1016/j.chb.2014.10.003
- Pardos, Z., & Baker, R. (2014). Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics*, 1(1), 107–128.
- Pardos, Z. A., Gowda, S. M., Ryan, S. J. D., & Heffernan, N. T. (2010). Ensembling Predictions of Student Post-Test Scores for an Intelligent Tutoring System. In *Proceedings of the 4th international conference on educational data mining* (pp. 189–198).
- Pedro, M. S., & Ocumpaugh, J. (2014). Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software. In *Proceedings of the 7th international conference on educational data mining* (pp. 276–279).
- Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6), 759–772. doi: 10.1109/TKDE.2008.138
- Rosé, C. P., & Siemens, G. (2014). Shared Task on Prediction of Dropout Over Time in Massively Open Online Courses. In *Conference on empirical methods in natural language processing workshop on modeling large scale social interaction in massively open online courses* (pp. 39–41).
- RUIPÉREZ-VALIENTE, J. A., MUÑOZ-MERINO, P. J., & DELGADO KLOOS, C. (2015). A Predictive Model of Learning Gains for a Video and Exercise Intensive Learning Environment. In *17th international conference on artificial intelligence in education (aied 2015)* (pp. 760–763).
- RUIPÉREZ-VALIENTE, J. A., MUÑOZ-MERINO, P. J., & DELGADO KLOOS, C. (2017). Detecting and Clustering Students by their Gamification Behavior with Badges: A Case Study in Engineering Education. *International Journal of Engineering Education*, 33(2-B), 816–830.

- Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Delgado Kloos, C., Niemann, K., & Scheffel, M. (2014). Do Optional Activities Matter in Virtual Learning Environments? In *Ninth european conference on technology enhanced learning* (pp. 331–344). Graz, Austria: Springer International Publishing. doi: 10.1007/978-3-319-11200-8_25
- Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Leony, D., & Delgado Kloos, C. (2015). ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. *Computers in Human Behavior*, 47(Learning Analytics, Educational Data Mining and data-driven Educational Decision Making), 139–148. doi: 10.1016/j.chb.2014.07.002
- Salehi, M., & Nakhai Kamalabadi, I. (2013). Hybrid recommendation approach for learning material based on sequential pattern of the accessed material and the learner's preference tree. *Knowledge-Based Systems*, 48, 57–69. doi: 10.1016/j.knsys.2013.04.012
- Santos, O. C., & Boticario, J. G. (2015). User-centred design and educational data mining support during the recommendations elicitation process in social online learning environments. *Expert Systems*, 32(2), 293–311. doi: 10.1111/exsy.12041
- Wang, T., & Mitrovic, A. (2002). Using neural networks to predict student's performance. In *Proceedings of 10th international conference on computers in education* (pp. 969–973). IEEE. doi: 10.1109/CIE.2002.1186127
- Yu, F. Y., & Wu, C. P. (2013). Predictive effects of online peer feedback types on performance quality. *Educational Technology and Society*, 16(1), 332–341.

AUTHOR BIOGRAPHY



José A. Ruipérez-Valiente completed his B.Eng. and M.Eng. in Telecommunications at Universidad Católica de San Antonio (UCAM) and Universidad Carlos III of Madrid (UC3M) respectively, graduating in both cases with the best academic transcript of the class. Afterwards, he completed his M.Sc. and Ph.D. in Telematics at UC3M while conducting research at Institute IMDEA Networks in the area of learning analytics and educational data mining. During this time, he completed two research stays of three months each, the first one at MIT and the second one at the University of Edinburgh. He has also held industry appointments at Vocento, Accenture and ExoClick. He has received several academic and research awards and has published more than 25 scientific publications in journals and conferences. Currently he is a postdoctoral associate at MIT where his research is focused on learning analytics and game-based assessment.



Pedro J. Muñoz-Merino received his Telecommunication Engineering degree in 2003 from the Polytechnic University of Valencia, and his PhD in Telematics Engineering in 2009 from the Universidad Carlos III de Madrid. He is a Visiting Associate Professor at the Universidad Carlos III de Madrid. He has done two long research stays: one in Ireland for more than 3 months at the Intel company in 2005, and another in Germany for more than 6 months at the Fraunhofer Institute of Technology in 2009-2010. He is author of more than 70 scientific publications and has participated in more than 20 research projects. He has been PC member of different conferences and invited as a speaker in different events in topics related to learning analytics and educational data mining. He is also an IEEE Senior Member from 2015.



Carlos Delgado Kloos received the PhD degree in Computer Science from the Technical University of Munich and in Telecommunications Engineering from the Technical University of Madrid. Since 1996, he is Full Professor of Telematics Engineering at the Universidad Carlos III de Madrid, where he is the Director of the online Master's program on "Management and Production of e-Learning", Holder of the UNESCO Chair on "Scalable Digital Education for All" and of the GAST research group. He is also Vice President for Strategy and Digital Education. He coordinates the eMadrid network on Educational Technology in the Region of Madrid. He is an IEEE Senior Member. His main interests are centered on educational technologies.

How to cite this article: J.A. Ruipérez-Valiente, P. J. Muñoz-Merino, and C. Delgado Kloos (2018), Improving the Prediction of Learning Outcomes in Educational Platforms including Higher Level Interaction Indicators, *Expert Systems*, 2017;00:1–6.