DOI: https://doi.org/10.1145/3462203.3475905

Data-Driven Performance Prediction in a Geometry Game Environment

Sofia Strukova
University of Murcia
Murcia, Spain
strukovas@um.es

José A. Ruipérez-Valiente
University of Murcia
Murcia, Spain
jruiperez@um.es

Félix Gómez Mármol
University of Murcia
Murcia, Spain
felixgm@um.es

ABSTRACT

The rapid technological evolution of the last years motivated students to develop competencies and capabilities that will prepare them for an unknown future of the 21st century. In this context, teachers intend to optimise the process of learning and make it more dynamic and exciting by introducing gamification. Thus, this paper focuses on a data-driven assessment of geometry competencies, which are essential for developing problem-solving and higher-order thinking skills. We explored them in the domain of knowledge inference, whose primary goal is to predict or measure the students' knowledge over questions as they interact with a learning platform at a specific time. Hence, the main goal of the current paper is to compare several well-known algorithms applied to the data of a geometry game named Shadowspect in order to predict students' performance in terms of classifier metrics such as Area Under Curve (AUC), accuracy, and F1 score. We found Elo to be the algorithm with the best prediction power. However, the rest of the algorithms also showed decent results, and, therefore, we can conclude that all the algorithms hold the potential to measure and estimate the actual knowledge of students. In turn, this means that they can be applied in formal education to improve teaching, learning, organisational efficiency and, as a consequence, this can serve as a basement for a change in the system.

CCS CONCEPTS

• **Applied computing** → **Education**; • **Information systems** → *Data mining*; • **Human-centered computing** → *Collaborative and social computing*.

KEYWORDS

Computational Social Science, Game-based Assessment, Knowledge Inference, Geometry Capabilities, Data mining.

ACM Reference Format:
Sofia Strukova, José A. Ruipérez-Valiente, and Félix Gómez Mármol. 2021. Data-Driven Performance Prediction in a Geometry Game Environment. In *Conference on Information Technology for Social Good (GoodIT '21), September 9–11, 2021, Roma, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3462203.3475905

GoodIT '21, September 9–11, 2021, Roma, Italy

ACM ISBN 978-1-4503-8478-0/21/09…$15.00
https://doi.org/10.1145/3462203.3475905

1 INTRODUCTION

Due to the rapid technological progress in the last years, we see a considerable change in the way of teaching, learning and personal development in general. Accordingly, we need to motivate students to develop competencies and capabilities that will prepare them for an unknown future. In this way, a lot of schools intend to improve the process of learning and make it more dynamic and exciting by introducing technology-mediated environments such as simulations, virtual reality or games. The latter advertised themselves as an excellent way to digitise and optimise the learning process [3] and showed significant evidence [2, 11] of benefits of their use for learning and assessment according to the student's ability to become competent in a specific field.

In the modern world, there is no doubt of the importance of geometry skills and spatial reasoning, which are essential human abilities contributing to mathematical proficiency. These skills can be crucial to functioning in twenty-first-century society, especially in careers associated with Science, Technology, Engineering and Mathematics (STEM). Concretely, Giofrè et al. explored the relationship across working memory, intelligence and geometry skills in children and concluded that working memory is strongly related to geometrical achievement irrespective of their intelligence [8]. While there are obvious benefits, geometry can lead to both anxiety in students and teaching difficulties in teachers. Gamification of the learning process might be a promising solution to these two issues, increasing the engagement and motivation of students [18].

Taking into account the emerging significance of geometry competencies for developing problem-solving and higher-order thinking skills and the proved reliability of game-based assessment (GBA), in our work, we will focus on Shadowspect[1], a game-based assessment tool that aims to provide metrics related to geometry content and other behavioural and cognitive constructs. In the context of games with the ability to educate, we are interested in knowledge inference, a sequence problem whose primary goal is to predict or measure the students' knowledge (or its absence) over questions as they are interacting with a learning platform at a specific time. It can be equivalent to monitoring Knowledge Components (KCs), which are associated with every problem-solving item. KCs were defined by Koedinger et al. as acquired units of cognitive function or structures that can be inferred from performance on a set of related tasks [9] and they generalise across terms for describing pieces of cognition or knowledge, including skills, concepts or facts. Through them, we can improve the knowledge of students, explore the influence of education and make automatic pedagogical decisions. Moreover, through knowledge inference algorithms, based on prediction, we measure how good the learner modelling

[1] https://shadowspect.org/

is – the estimation of actual knowledge of students. Ultimately, this can be used as part of the formative assessment process, where the data generated by learners hold the potential for being used as part of such formative assessment process and for *adaptive learning*, whose goal is to address the unique needs of each user [10]. In the paper at hand, we will perform comparative research of BKT, PFA and Elo algorithms for predicting the performance of learners in the context of the Shadowspect game. To the best of our knowledge, the above-mentioned knowledge inference approaches have not yet been applied in GBA. We believe this is an important novelty that could transform the educational process significantly.

The remainder of this paper is structured as follows. In Section 2, we focus on the background of our study and related works. In Section 3, we present the research methodology. Our findings are outlined in Section 4. Finally, we draw our conclusions and future research directions in Section 5.

2 BACKGROUND

As a general rule, it is not a trivial task to measure or predict the capabilities of users. There are two main reasons why a student's performance in a specific task attempt might not mean that the student has the skill: 1) the student can **slip**, which means not to demonstrate the skill *despite* having it, and 2) the student can **guess**, which means to demonstrate the skill *without* having it. Moreover, we cannot directly estimate these skills. Even so, we can measure knowledge inference by looking at the performance of the student over time. In this way, there is a wide variety of methods that aim to measure the existing knowledge and forecast the future outputs of users. The first proposed method for observing students' past successes and failures was Bayesian Knowledge Tracing (BKT) [4], which employs a two-state dynamic Bayesian network estimating the latent cognitive state from students' performance where each KC is either learned or unlearned. An alternative approach for performance prediction is Performance Factors Analysis (PFA) [13], which uses a logistic regression equation that models changes in performance in terms of the number of student successes and failures that have occurred for each skill [19]. Finally, another approach is the Elo rating system (named after its creator Arpad Elo) [6, 15] – a variant of Item Response Theory (IRT), whose classical approach has some key limitations. The main idea of the Elo algorithm is to continually estimate the difficulty of an item and the ability of a student, updating both of them every time a student encounters an item.

The task of solving the knowledge inference problem attracted many researchers. There are several authors who conducted surveys comparing different variations of the approaches mentioned above [1]. These works served as a base for others to conduct experiments on real-world data sets. For example, Gervet et al. [7] analysed the performance of various algorithms such as Deep Knowledge Tracing (DKT) [16], IRT, PFA, and BKT, amongst others, exhibiting two main advantages with respect to other articles: 1) it explored a wide variety of methods to predict the learner performance and 2) the efficiency of the algorithms as mentioned earlier was proved on nine real-world data sets with different characteristics, i.e., the number of items and KCs they cover, the number of learners or total interactions they contain. The authors concluded that DKT leads on data sets of large size or where precise temporal information matters most. In contrast, others can perform better on data sets of moderate size or containing a vast number of interactions per student.

From the works mentioned above, we can observe the fact that the research covered by these areas is currently and constantly increasing. On the other hand, the above-mentioned articles worked in Intelligent Tutoring Systems (ITS), where it is easier to model student learning because they have clearly defined tasks. Modelling learning in games is more challenging because they are more open environments where students should keep a friendly and motivating atmosphere all the time. In our case study, the game-based assessment tool Shadowspect was previously designed for the very purpose of measuring geometry content standards so that teachers can use it in their core geometry curriculum.

3 METHODOLOGY

In this section, we will describe our research goals (RGs). Next, we will characterise the context of the geometry game environment employed in this work. Finally, we will give details of the adjustment of the Shadowspects' data for BKT, PFA and Elo algorithms, discuss each of them and the metrics which will explain their performance.

3.1 Research goals

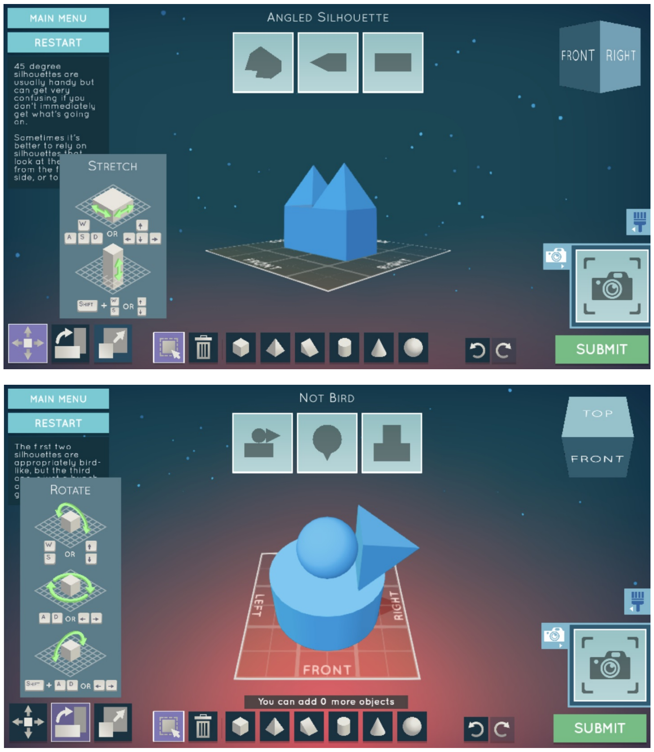
After examining state of the art regarding the knowledge inference problem, application of its algorithms and the existing related works, we stated the following research goals:

- **RG1**. To compare BKT, PFA and Elo algorithms applied to the data of a geometry game named Shadowspect in order to predict students' performance in terms of classifier metrics such as AUC, accuracy and F1 score.
- **RG2**. To analyse if the algorithms outperform in predicting performance in any particular KC.

3.2 Context of the game environment

The game environment Shadowspect was developed at the Massachusetts Institute of Technology (MIT) Playful Journey Lab, and it has clearly defined goals, rules, obstacles for the players to overcome and provides only intrinsic rewards [17]. In the version of Shadowspect (see Fig. 1) that we used in this case study, there are nine tutorial levels (teaching the basic functionality of the game, i.e., how to build different primitives, scale and rotate them), nine intermediate levels (giving students more freedom so they will not receive much help to solve the puzzles) and 12 advanced levels (challenging the students who already proved to gain experience). When students begin a puzzle, they receive a set of silhouettes from different views that represent the figure they need to create by using other primitive shapes (i.e., cubes, pyramids, ramps, cylinders, cones and spheres), which can be scaled, moved and rotated. Moreover, the students can move the camera to see the figure they are building from different perspectives and then use the 'Snapshot' functionality to generate the silhouette and see how close they are to the specified goal. Finally, the students can submit the puzzle, and the system will evaluate the solution and provide feedback.

The KCs in the game environment Shadowspect are the skills needed to complete a puzzle successfully. Across Shadowspect,



45 degree silhouettes are usually handy but can get very confusing if you don't immediately get what's going on.

Sometimes it's better to rely on silhouettes that look at the from the f side, or to

The f rst two silhouettes are appropriately bird-like, but the third

Figure 1: Two puzzle examples in Shadowspect

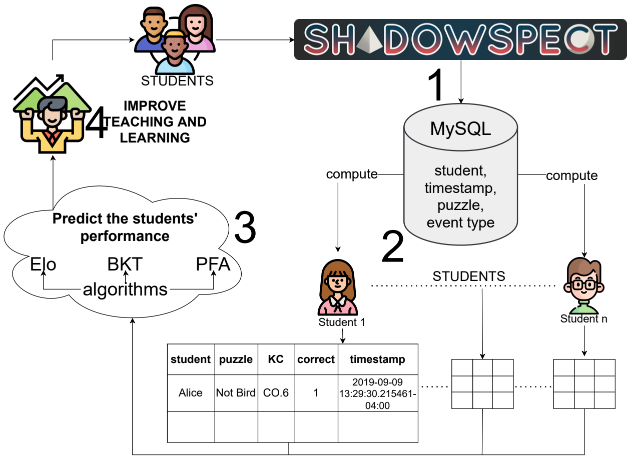
experts defined four main KCs, and most of the puzzles have the representation of three (GMD.4, CO.5 and CO.6) of them:

- **MG.1**: Use geometric shapes, their measurements and their properties to describe objects.
- **GMD.4**: Identify the shapes of the two-dimensional cross sections of the three-dimensional objects and identify the three-dimensional objects generated by the rotations of the two-dimensional objects.
- **CO.5**: Given a geometrical figure and a rotation, reflection or translation, draw the transformed figure using, for example, graph paper, tracing paper or geometry software. Specify a sequence of transformations that will take one given figure to another.
- **CO.6**: Use geometric descriptions of rigid movements to transform figures and predict the effect of a given rigid movement on a given figure; in the case of two figures, use the definition of congruence in terms of rigid movements to decide if they are congruent.

Both puzzles represented in Fig. 1 assume that to solve them, the student must have the following KCs: GMD.4, CO.5 and CO.6. It implies that each KC means to have the same proportion (33%) because there are three KCs in one puzzle. Moreover, most of the puzzles (90%) reflect the same idea of requiring the same three KCs while the rest of the puzzles request one more KC, namely MG.1. In practice, it is not precise because one KC might play a dominant role. For the reason that it is a complex task to implement the algorithms taking into account the correct weights of each KC in a puzzle, in our work, we will assume that all the present KCs have the same weight.

3.3 Adjusting Shadowspects' data for the algorithms

For this paper, we used the data from 322 different students, which were collected as part of the assessment machinery development. The complete data collection recorded in an input experiment document includes around 428,000 events (an average of 1,320 events per user). Students were active in the game environment for 260 hours (an average of 0.82 active hours per student), and students solved a total of 3,802 puzzles (an average of 13 puzzles per student). All student interactions with the game were stored in a MySQL database, and we did not collect any identifiable or personal data from the users except a nickname provided by themselves.



student	puzzle	KC	correct	timestamp
Alice	Not Bird	CO.6	1	2019-09-09 13:29:30.215461-04:00

Figure 2: Overview of methodology to predict the students' performance

To be consistent with the common steps, we made the same data-related assumptions for all the algorithms. The entire methodology process to predict the students' performance is represented in Figure 2. In the first step, we iterated through the input experiment document, identifying and storing the types of events (e.g., start a game, complete a puzzle, create, move, rotate, scale or delete a shape, exit to the menu, etc.) aligned with each user, the timestamp and the name of the puzzle in which the events occurred. As the events of the data are ordered according to time, and the events of different students are interspersed, we made the separation between the events of each student. Accordingly, in the second step, we computed if the student was correct or not in his attempt to solve the puzzle. At this point, we saw that our data set was imbalanced because we had many more records of students solving the puzzles than being incorrect in their attempts. In this way, it is crucial to describe the puzzle-related assumptions. Firstly, if a student made one submission of his solution to the puzzle, we considered it as one attempt. Moreover, if a student already solved the puzzle and made another attempt to complete it, we discarded the latter. We also did not count as an attempt the situations when the student made no submissions.

In the third step, we implemented BKT, PFA and Elo algorithms for our case study in order to predict the student' performance based on the current modelling of each student. Finally, with this value in mind, the teacher can know the learner's ability before making a formal assessment, intervene to see the cause of strange behaviour and help the student to improve. Through these measures, we seek to collaborate both in the students' evaluation and in adaptive learning. In addition to the indicators of competence and difficulty, we obtained a prediction model with the probabilities of future success.

3.4 Algorithms

In this section, we describe the key approaches for knowledge inference, namely: **BKT**, **PFA** and **Elo** algorithms.

3.4.1 Bayesian Knowledge Tracing (BKT). BKT [4] estimates the students' knowledge from their observed actions – the history of performance with that skill. This algorithm maintains a continuous evaluation of the probability that a student currently knows each skill, updating that estimated value based on the student's behaviour [5]. In this algorithm, only the first attempt on each item matters and learning is modelled by a discrete transition from an unknown to a known state. A fundamental assumption is that the student does not forget a skill once he knows it.

The advantage of BKT is that it is easy to interpret the parameters as well as their effects on performance in the model. The standard BKT model is using the following probabilities:

- $p(L_0)$ - the probability that the student has prior knowledge meaning that he knows a KC before practising on any items associated with the KC;
- $p(T)$ - the probability of learning, meaning that the student will learn a KC by practising;
- $p(G)$ - probability that the student will guess the item correctly;
- $p(S)$ - probability that the student will slip.

Based on these parameters, the inference is made about the student's probability of knowledge at time opportunity n, $p(Ln)$. The parameters and inferred probability of knowledge can also be used to predict the correctness of a student response. The following equations are used to predict students' knowledge from behaviour in BKT:

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1}) * (1 - P(S))}{P(L_{n-1}) * (1 - P(S)) + (1 - P(L_{n-1})) * P(G)} \tag{1}$$

$$P(L_n|Action_n) = P(L_{n-1}|Action_n) + ((1 - P(L_{n-1}|Action_n)) * P(T) \tag{2}$$

3.4.2 Performance Factors Analysis (PFA). The goal of the PFA model is to measure how much skill a student has while learning. It has considerable power to fit data and provides the adaptive flexibility to create the model overlay to be used adaptively by a tutor [13]. In the standard PFA model, the data about learner performance are used to compute a skill estimate, and this estimate is then transformed using a logistic function into the estimate of the probability of a correct answer. In this way, the model is using the following parameters:

- β - the easiness of the KC;
- s - the prior successes for the KC of the student;
- f - the prior failures for the KC of the student;
- γ and ρ - success learning rate and failure learning rate of each skill, respectively.

Equation 3 reveals how to compute the probability $P(m)$ that the learner i will get the item k correct where m is a logit value representing the accumulated learning for student i (ability captured by γ parameter) using a KC j.

$$P(m) = \frac{1}{1 + e^{-m}} \tag{3}$$

$$m(i, k \epsilon Items, s, f) = \beta_k + \gamma s_i + \rho f_i \tag{4}$$

We find several strengths to be considered for PFA in our environment: it softens the impact of incorrect answers so that the created model is more realistic and does not modify as much in the face of error. Besides, it has an essential advantage over BKT because it does not consider errors in the exercises as decisive and implies a more gradual modification. Despite the benefits it brings, it is a fairly complex algorithm to implement since it takes into account numerous factors that make it difficult to adapt (e.g., the difficulty parameter). Moreover, PFA does not take into account the order in which past successes and failures occurred in.

3.4.3 Elo rating system. Elo [15] is a skill calculation system used, for example, in chess tournaments. It was developed for the purpose of measuring players' strength, but it also was applied in the context of educational research and was used for measuring both learner ability and task difficulty [12]. Its basic principle is as follows: a score is assigned to each player, and then this score is updated after each game proportionally to how surprising the result of the game was (if a weak player beats a strong one, the results were unexpected and therefore the update is big). First, we must obtain the probability that a student answers correctly a question by using a logistic function with both the competence of the student θ_s and the difficulty of the question d_i while the correctness of an answer of a student on an item is $correct_{si}$ ϵ {0,1}:

$$P(correct_{si} = 1) = \frac{1}{(1 + e^{-(\theta_s - d_i)})} \tag{5}$$

Next, we calculate the probability of each student-question confrontation. Initial values of θ_s and d_i parameters are set to 0. The value of the constant K determines the behaviour of the system (i.e., if K is small, the estimation converges too slowly). The following equations represent updates for both the competence of the student and the difficulty of the puzzle:

$$\theta_s = \theta_s + K * (correct_{si} - P(correct_{si} = 1))) \tag{6}$$

$$d_i = d_i + K * (P(correct_{si} = 1)) - correct_{si}) \tag{7}$$

The implementation and adaptation of the Elo algorithm to the data are not complicated. The algorithm has few adjustment parameters, and it is also computationally very simple and fast [20].

Moreover, it competes in performance with other much more complex algorithms and can be implemented in almost any type of data, being able to modify it in a straightforward way.

3.5 Classifier Metrics

We will be using each algorithm to obtain a standardised numerical value between 0 and 1 for the geometry capabilities according to each KC based on the history of each student's interactions with the activity. After exploring the work performed by Pelánek [14], who made an overview of all commonly used metrics and discussed their properties, advantages and disadvantages applied to educational data mining, we decided to rely on the following metrics for comparing the applied algorithms between each other keeping in mind that our data set is imbalanced:

- **Accuracy** - is the total percentage of correctly classified elements. In other words, accuracy looks at fractions of correctly assigned positive and negative classes.
- **AUC** - is a more comprehensive measure of how good the classifier is at distinguishing between classes. In other words, AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example. This metric is representative but not discriminative. The higher the AUC, the better the model is at correct predictions.
- **F1 score** - is a measure of a test's accuracy, which is calculated based on the precision and recall of the test. The recall is the number of true positive results divided by the number of all samples that should have been identified as positive. The classical counterpart to recall is precision which is the number of true positive results divided by the number of all positive results, including those not identified correctly.

For our particular case study, we first chose a more traditional accuracy metric. The overall accuracy depends on the ability of the classifier to rank patterns and also on its power to select a threshold in the ranking used to assign patterns to the positive class if above the threshold and to the negative class if below. The classifier with the higher AUC metric is likely to also have higher overall accuracy as the ranking of patterns is beneficial to both AUC and overall accuracy. However, if one classifier ranks patterns well but selects the threshold badly, it can have a high AUC but a poor overall accuracy. On the other hand, both metrics look at fractions of correctly assigned positive and negative classes. It means that if our problem is highly imbalanced, we can get high scores by simply predicting that all observations belong to the majority class. In this way, the last metric we decided to include is the F1 score that works well with such cases. We believe that with these three metrics, we can reasonably conclude the performance of the selected algorithms taking into consideration all the advantages and disadvantages of the metrics.

4 RESULTS

In this section, we will highlight our findings following the stated RGs. First, we will discuss the results obtained by building each algorithm and comparing them with the use of selected metrics. Next, we will analyse the performance of each algorithm in accordance with each KC.

4.1 Algorithms comparison (RG1)

	AUC	Accuracy	F1 score
BKT	0.79	0.86	0.92
PFA	0.84	0.86	0.92
Elo	0.85	0.94	0.97

Table 1: Comparison of BKT, PFA and Elo algorithms by AUC, accuracy and F1 score metrics

From Table 1, we can observe the fact that all the algorithms show adequate results in the stated metrics. The accuracy of BKT and PFA algorithms is identical, while the Elo algorithm outperforms them by 8%. We also see that no significant difference was detected in the AUC metric in the comparison of PFA and Elo models, but in this case, BKT slightly underperforms them. With these values in mind, we conclude that the models have a very high precision indicating that the adjustment was carried out correctly. However, we should always consider an imbalance when looking at the accuracy and the AUC metrics. Finally, since we have a skewed sample distribution, in Section 3.5 we decided to use both precision and recall. The results of the F1 score metric reveal that all the algorithms indicate outstanding precision and recall. Therefore, we can deduce that BKT, PFA and Elo algorithms are precise and robust.

The results reveal that all the algorithms show decent outputs. On average, we see that BKT performs slightly more inferior but still, we observe reasonable results considering our case study. While PFA also performs sufficiently well, we see that the Elo algorithm outperforms others in the most critical metrics AUC and F1 score. Therefore, we conclude that the most predictive model is Elo, which outperforms the overall accuracy of BKT and PFA by 8% and F1 score by 5%.

4.2 Algorithms performance per KC (RG2)

For further analysis, we computed the accuracy metric per KC. The according results are represented in Figure 3. First of all, the analysis confirmed our findings stating that Elo outperforms the rest of the algorithms.

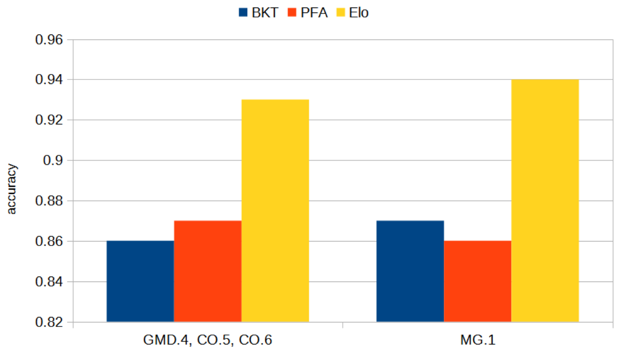


Figure 3: Accuracy of each algorithm per KCs

Secondly, we see that we obtain the same results for all the metrics regarding three KCs (GMD.4, CO.5 and CO.6). This could be predicted at the point when we saw how KCs match with each other. As said in Section 3.2, most of the puzzles have the representation of these KCs while others additionally have the fourth KC named MG.1. Accordingly, if a puzzle requires the proper use of KC CO.5, it means that the student must also apply CO.6 and GMD.4. Since in this version of the implementation, we have neither dominant KCs neither according weights assigned to them, the metrics would show the same outcome. Thus, for more precise results, there is a need of adding the weights to KCs in future work.

5 CONCLUSIONS AND FUTURE DIRECTIONS

Understanding learners and their contexts has undoubtedly become one of the most promising educational research topics of the past decade. Accordingly, every year there are more novel solutions to promote various educational settings and motivate students. In this way, gamification proved to be an important way of engaging students. This work represents a novel analysis and comparison of three algorithms, namely, BKT, PFA and Elo, applied to a geometry game environment in order to predict the learning performance of its users. We measured the efficiency of the algorithms as mentioned earlier by examining the following metrics: AUC, accuracy and F1 score. We found Elo to be the algorithm with the best prediction power. However, the rest of the algorithms also showed decent results and, therefore, we can conclude that they all hold the potential to measure and estimate the actual knowledge of students. In turn, this means that all the three analysed algorithms are suitable for the application in formal education to improve teaching, learning, organisational efficiency and, as a consequence, this can serve as a basement for a change in the system. On the other hand, we are confident that this work could motivate teachers and students to use gamification for the learning process. Moreover, this experience could also be transferred into not formal educational settings with new innovative products.

As far as we are aware, this is the first time the research was conducted on applying and comparing the above-mentioned knowledge inference models in GBA for predicting the learners' performance. Besides, there are several possible extensions to this research. Our future work will focus on implementing the similar BKT, PFA and Elo algorithms but considering the weights of KCs in each puzzle. In this way, the results will be more precise what is essential for making the difficulty and competencies calculations. Moreover, we will intend to explore and apply other models, i.e., Deep Knowledge Tracing.

ACKNOWLEDGEMENTS

This study was partially funded by the Spanish Government grants FJCI-2017-34926 and RYC-2015-18210, co-funded by the European Social Fund, as well as by the COBRA project (10032/20/0035/00), granted by the Spanish Ministry of Defence.

REFERENCES

[1] Ali Alkhatlan and Jugal Kalita. 2018. Intelligent tutoring systems: A comprehensive historical survey with recent developments. *arXiv preprint arXiv:1812.09628* (2018).

[2] Stacey Brull and Susan Finlayson. 2016. Importance of gamification in increasing learning. *The Journal of Continuing Education in Nursing* 47, 8 (2016), 372–375.

[3] Fu Chen, Ying Cui, and Man-Wai Chu. 2020. Utilizing Game Analytics to Inform and Validate Digital Game-based Assessment with Evidence-centered Game Design: A Case Study. *International Journal of Artificial Intelligence in Education* 30, 3 (2020), 481–503.

[4] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[5] Ryan SJ d Baker, Albert T Corbett, and Vincent Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *International conference on intelligent tutoring systems*. Springer, 406–415.

[6] Arpad E Elo. 1978. *The rating of chessplayers, past and present.* Arco Pub.

[7] Theophile Gervet, Ken Koedinger, Jeff Schneider, Tom Mitchell, et al. 2020. When is Deep Learning the Best Approach to Knowledge Tracing? *JEDM| Journal of Educational Data Mining* 12, 3 (2020), 31–54.

[8] David Giofrè, Irene Cristina Mammarella, and Cesare Cornoldi. 2014. The relationship among geometry, working memory, and intelligence in children. *Journal of Experimental Child Psychology* 123 (2014), 112–128.

[9] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.

[10] Min Liu, Emily McKelroy, Stephanie B Corliss, and Jamison Carrigan. 2017. Investigating the effect of an adaptive learning intervention on students' learning. *Educational technology research and development* 65, 6 (2017), 1605–1625.

[11] Pedro A Martınez, Manuel J Gómez, Jose A Ruipérez-Valiente, Gregorio Martınez Pérez, and Yoon Jeon Kim. 2020. Visualizing Educational Game Data: A Case Study of Visualizations to Support Teachers. (2020).

[12] Maciej Pankiewicz and Maricn Bator. 2019. Elo Rating Algorithm for the Purpose of Measuring Task Difficulty in Online Learning Environments. *e-mentor* 5 (82) (2019), 43–51.

[13] Phil Pavlik Jr, Hao Cen, and Kenneth Koedinger. 2009. Performance Factors Analysis - A New Alternative to Knowledge Tracing. *Frontiers in Artificial Intelligence and Applications* 200, 531–538. https://doi.org/10.3233/978-1-60750-028-5-531

[14] Radek Pelánek. 2015. Metrics for Evaluation of Student Models. *Journal of Educational Data Mining* 7, 2 (2015), 1–19.

[15] Radek Pelánek. 2016. Applications of the Elo rating system in adaptive educational systems. *Computers & Education* 98 (2016), 169–179.

[16] Chris Piech, Jonathan Spencer, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *arXiv preprint arXiv:1506.05908* (2015).

[17] José A Ruipérez-Valiente and Yoon Jeon Kim. 2020. Effects of solo vs. collaborative play in a digital learning game on geometry: Results from a K12 experiment. *Computers & Education* 159 (2020), 104008.

[18] José A Ruipérez-Valiente, Pedro J Muñoz-Merino, and Carlos Delgado Kloos. 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 (2017), 816–830.

[19] Richard Scruggs, Ryan S. Baker, and Bruce M. McLaren. 2020. Extending Deep Knowledge Tracing: Inferring Interpretable Knowledge and Predicting Post-System Performance. arXiv:1910.12597 [cs.CY]

[20] Angela Verschoor, Stéphanie Berger, Urs Moser, and Frans Kleintjes. 2019. On-the-Fly Calibration in Computerized Adaptive Testing. In *Theoretical and Practical Advances in Computer-based Educational Measurement.* Springer, Cham, 307–323.