Strukova, S., J. A. Ruipérez-Valiente, and Gómez Mármol, F., "Adapting Knowledge Inference Algorithms to Measure Geometry Competencies through a Puzzle Game," in ACM Transactions on Knowledge Discovery from Data, 18(1), Article No.: 21, Pages 1 - 23 https://doi.org/10.1145/3614436.

https://dl.acm.org/doi/abs/10.1145/3614436

Adapting Knowledge Inference Algorithms to Measure Geometry Competencies through a Puzzle Game

SOFIA STRUKOVA, Department of Information and Communication Engineering, University of Murcia, Spain
JOSÉ A. RUIPÉREZ-VALIENTE, Department of Information and Communication Engineering, University of Murcia, Spain
FÉLIX GÓMEZ MÁRMOL, Department of Information and Communication Engineering, University of Murcia, Spain

The rapid technological evolution of the last years has motivated students to develop capabilities that will prepare them for an unknown future in the 21st century. In this context, many teachers intend to optimise the learning process, making it more dynamic and exciting through the introduction of 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. Our main goal is to adapt, evaluate and compare Bayesian Knowledge Tracing (BKT), Performance Factor Analysis (PFA), Elo and Deep Knowledge Tracing (DKT) algorithms applied to the data of a geometry game named Shadowspect, in order to predict students' performance by means of several classifier metrics. We analysed two algorithmic configurations, with and without prioritisation of Knowledge Components (KCs) – the skills needed to complete a puzzle successfully, and we found Elo to be the algorithm with the best prediction power with the ability to model the real knowledge of students. However, the best results are achieved without KCs because it is a challenging task to differentiate between KCs effectively in game environments. Our results prove that the above-mentioned algorithms can be applied in formal education to improve teaching, learning, and organisational efficiency.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Computational Social Science, Data-driven Evaluation, Data Mining, Competencies, Capabilities

ACM Reference Format:
Sofia Strukova, José A. Ruipérez-Valiente, and Félix Gómez Mármol. 2022. Adapting Knowledge Inference Algorithms to Measure Geometry Competencies through a Puzzle Game. 1, 1 (January 2022), 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent years have been marked by rapid technological advancements, leading to a considerable change in the way of teaching, learning and personal development in general. Accordingly, it has become necessary to motivate students to develop competencies and capabilities that prepare them for an unknown future and contemporary challenges [48]. In response to this need, many schools intend to improve the learning process by making it more dynamic and exciting through the introduction of technology-mediated environments such as simulations, virtual reality or games. The latter are a good way to digitise and optimise the learning process [8] and has demonstrated significant benefits

---

Authors' addresses: Sofia Strukova, Department of Information and Communication Engineering, University of Murcia, Murcia, Spain, strukovas@um.es; José A. Ruipérez-Valiente, Department of Information and Communication Engineering, University of Murcia, Murcia, Spain, jruiperez@um.es; Félix Gómez Mármol, Department of Information and Communication Engineering, University of Murcia, Murcia, Spain, felixgm@um.es.

Manuscript submitted to ACM

for both learning and assessment, given the student's ability to become competent in a specific field [6, 30]. Within these game-based settings, gamification assessment plays a critical role. It involves evaluating student learning and performance within these games, often incorporating various elements to motivate students. Gamification assessment also provides immediate feedback and clearly defined goals, enabling students to track and improve their learning progress [26]. Besides, it allows for adaptive difficulty levels and personalised learning pathways, providing valuable insights into students' performance and behaviour within the game. This not only fosters competency in specific fields but also promotes the development of higher-order thinking skills.

In this purview, one of the most highly valued capabilities in the modern world is mathematical proficiency. Within mathematics, there is no doubt about the importance of geometry skills and spatial reasoning because they are essential human abilities that contribute to mathematical expertise. These skills can be crucial to functioning in the twenty-first-century society, especially in careers associated with Science, Technology, Engineering and Mathematics (STEM). More specifically, Giofrè et al. [18] 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. While there are obvious benefits, geometry can lead to both anxiety in students and teaching difficulties in teachers. The gamification of the learning process or game-based settings might be a promising solution to these two issues, increasing the engagement and motivation of students [43]. They both can include similar gaming elements such as points, badges, and challenges. However, the main difference in the use of elements and features between gamification and game-based settings is that gamification uses game-like elements to enhance an existing process, while game-based settings use a fully-fledged game to teach or reinforce a specific skill or objective.

In our work, taking into account the emerging significance of geometry competencies for developing problem-solving and higher-order thinking skills and the proven reliability of game-based assessment (GBA) [22], we will focus on Shadowspect[1], a GBA tool developed by the Massachusetts Institute of Technology (MIT) that aims to provide metrics related to geometry content and other behavioural and cognitive constructs. In the context of games for educational purposes, we are interested in knowledge inference, a sequence problem whose primary goal is to predict or model the students' knowledge (or gaps in their knowledge) 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. [23] as acquired units of cognitive function or structures that can be inferred from performance on a set of related tasks, and they generalise across terms for describing pieces of cognition or knowledge, including skills, concepts or facts. Through them, we can improve students' knowledge, explore the influence of education and make automatic pedagogical decisions. Moreover, through knowledge inference algorithms and based on the outcome prediction of the tasks attempted, we measure how good the learner modelling is – the estimation of actual latent knowledge of students. Ultimately, this can be used as part of the formative assessment process, where data generated by learners can potentially guide *adaptive learning*, whose goal is to address the unique needs of each user [27].

Notwithstanding, digital GBAs yield student process data that are much more difficult to analyse than traditional assessments [11]. Moreover, several studies prove that knowledge tracing algorithms can be an effective tool for improving learning outcomes in games by providing personalised support and feedback, and identifying areas of difficulty for targeted intervention. Among them, Nagatani et al. extended the deep knowledge tracing model in order to model and predict a student's knowledge by considering their forgetting behaviour [33] and Long et al. with the usage

---

[1] https://shadowspect.org/

of knowledge tracing estimated the students' cognition on the question before response prediction and assessed their knowledge acquisition sensitivity on the questions before updating the knowledge state [29]. However, to the best of our knowledge, the above-mentioned knowledge inference approaches have not yet been applied and compared in GBA to predict the learners' performance [19]. Therefore, in the paper at hand, we will perform adaptive and comparative research of Bayesian Knowledge Tracing (BKT), Performance Factors Analysis (PFA), Elo and Deep Knowledge Tracing (DKT) algorithms for predicting learners' performance in the context of the Shadowspect game. We chose these algorithms because they represent the most common ones showing the best results applied to similar problems in non-game contexts. To the best of our knowledge, this is the first study that will adapt and evaluate all these algorithms to GBA. Additionally, in our work, we question the significance of KCs in GBA following our previous work [47]. Previously, we applied the BKT, PFA and Elo algorithms to Shadowspect to predict the performance of students not separating or prioritising KCs. Now, we will perform additional experiments in order to reflect on the improvement of algorithms' performance. We presuppose that the use of KCs can help to unpack more complex constructs within individual KCs that will provide the opportunity for teachers to understand better the behaviour of students and give them more personalised feedback. In this way, we intend to suggest new ways of evaluating both the performance of students and the effectiveness of adaptive learning approaches. 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, while we extend the results in Section 5. Finally, we draw our conclusions and future research directions in Section 6.

2 BACKGROUND

2.1 Knowledge inference

As a general rule, it is not a trivial task to measure or predict the latent knowledge of learners. 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 an extensive 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 BKT [9]. BKT employs a two-state dynamic Bayesian network to estimate the latent cognitive state from students' performance where each KC is either learned or unlearned. An alternative approach for performance prediction is PFA [36], 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 [46]. Another approach is the Elo rating system (named after its creator Arpad Elo) [16, 38] – a version of Item Response Theory (IRT), whose classical approach has some fundamental limitations. In our work, we will not use the original algorithm but its variant design to perform learner modelling. 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. Finally, Deep Knowledge Tracing (DKT) [40] predicts performance on future items within a system based on Recurrent Neural Networks (RNNs) that are 'deep' in time to the task of knowledge tracing. In the following subsection, we will analyse the works examining and/or applying the above-mentioned knowledge inference algorithms in order to build properly the methodology of the paper at hand. A complete overview of all the existing knowledge tracing algorithms, their categorisation from a technical perspective and research gaps in their application can be found in a survey carried out by Liu et al. [28].

2.2 Related work

The task of solving the knowledge inference problem has attracted a lot of researchers. Several authors conducted surveys comparing different variations of the above-mentioned approaches [2, 39]. These works served as a base for others to conduct experiments on real-world data sets. For example, Sahebi et al. [44] explored several log-driven approaches predicting student performance in solving exercises on a data set collected from the online self-assessment system, which provides parameterised questions for learning Java programming. After comparing BKT, PFA, the advanced collaborative filtering approach (three dimensional and four-dimensional tensor factorisation) [24], and Feature-Aware Student Knowledge Tracing [20], the authors concluded that the last two approaches produce better performance by taking into account the knowledge structure. Also, Gervet et al. [17] analysed the performance of various algorithms such as DKT [40], IRT, PFA, and BKT, amongst others, exhibiting two main advantages with respect to other articles: 1) they explored a wide variety of methods to predict the learner performance, and 2) the efficiency of the algorithms as mentioned above 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 large data, or where precise temporal information matters most. In contrast, other algorithms can perform better on data sets of moderate size, or containing a vast number of interactions per student. This also happens in more typical applications of deep learning – for example, Dutta and Gros classified medical images and depending on the training size and network layout used, the accuracy changed from 74.1% for the smallest training size and increasing to 92.3% for the biggest training size [15].

From the works mentioned above, we can observe that the research covered by these areas is currently and constantly increasing. Even so, these articles worked in Intelligent Tutoring Systems (ITSs), where it is easier to model student learning because they have clearly defined tasks. ITS offers individual tutoring benefits automatically and autonomously, making each user progress at their own pace [14]. One of the fundamental principles of the ITS states that a learner borrows the needed information from worked examples and connects the new information with the prior knowledge [32]. In contrast, 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. Accordingly, we found a few works applying the algorithms mentioned above to GBA. For example, [11] applied BKT and Dynamic Bayesian Networks (DBNs) for analysing students' response process data from an interactive GBA. This game measures if students can interpret weather phenomena and the related human activity in a correct way. The authors found that BKT and DBNs proved to be valuable and informative for analysing the process data for their particular GBA case study, considering that they both allow tracking of the state of students mastery of knowledge and skills, which are changing over the game. Based on the findings of this study, we will go beyond by adapting more knowledge inference algorithms and doing experiments on prioritising certain KCs.

After exploring related works, we are confident that the GBA tool Shadowspect, which was previously designed for the very purpose of measuring geometry content standards so that teachers can use it in their core geometry curriculum, can also benefit from the use of the above-mentioned processes. Moreover, as far as we know, this is the first study that will adapt and apply a wide range of knowledge inference algorithms to GBA.

3 METHODOLOGY

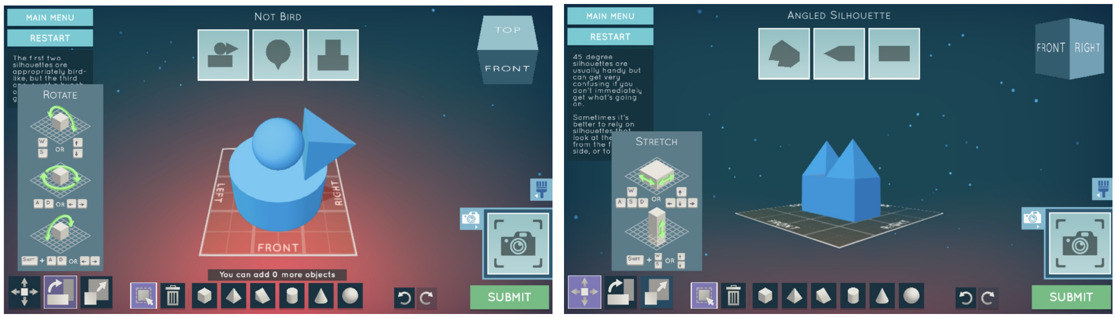
In this section, we will describe our research goals (RGs), we will characterise the context of the geometry game environment employed in this work, and, finally, we will give details of the BKT, PFA, Elo and DKT algorithms implementation that we performed, discussing each of them and the metrics that we used to measure their performance.

3.1 Research goals

After examining the state of the art regarding the knowledge inference problem applied to educational settings and to games, application of its algorithms and the existing related works, we stated the following research goals (RGs):

- **RG1**. To adapt and compare the effectiveness of four different knowledge inference algorithms – Bayesian Knowledge Tracing (BKT), Performance Factor Analysis (PFA), Elo, and Deep Knowledge Tracing (DKT) – within the context of a geometry game called Shadowspect. The efficacy of these algorithms will be measured by their capacity to predict students' performance and accurately model their knowledge, using classifier metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC), accuracy, and F1 score. This objective further includes an examination of the impact of Knowledge Components (KCs) – the skills required to solve a puzzle successfully – on the performance of learner modelling. We aim to explore how the inclusion or exclusion of KCs affects the performance of the aforementioned algorithms in predicting and modelling student knowledge.
- **RG2**. After identifying the algorithm with the highest performance from the RG1, our second goal is to perform a thorough evaluation of this algorithm. The evaluation process involves computing metrics such as accuracy per puzzle, accuracy per user, puzzle difficulty, and individual student competence. These metrics will provide a comprehensive understanding of students' behaviours, which can subsequently inform the development of personalised feedback mechanisms in game-based learning environments.

3.2 Context of the game environment



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

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

Fig. 1. Two puzzle examples in Shadowspect

The game environment Shadowspect was developed at the MIT Playful Journey Lab, and its overall objective is to solve geometry puzzles. It has clearly defined goals, rules, obstacles for the players to overcome and it provides only intrinsic rewards [42]. In the version of Shadowspect (see Figure 1) that we used for 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 do 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 representing 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. Based on the proven importance of geometry skills, the ability of the Shadowspect game environment to measure them and the overall satisfaction already reported by students while playing, we are confident that choosing this puzzle game for our experiments has great potential.

Shadowspect was specifically chosen for this study due to several reasons that made it an ideal platform for this kind of investigation. Firstly, Shadowspect is a highly engaging and interactive geometry puzzle game, which incorporates a robust range of geometry-related tasks and has clearly defined goals and rules, making it an excellent tool for learning and assessment [42]. More importantly, Shadowspect presents tasks associated with multiple KCs. The game's tasks offer a lot of opportunities for assessing students' understanding and application of these geometry concepts. This diverse range of KCs, in turn, contributes to the rich data set that allowed us to effectively adapt, evaluate and compare the knowledge inference algorithms in our study.

Furthermore, the choice of Shadowspect highlights the generalizability of our study. While Shadowspect serves as an ideal case study environment due to its comprehensive integration of geometric KCs and gameplay design that encourages learning, our approach is not limited to this specific game. In fact, our methodology can be extended to any educational game that embodies certain characteristics, such as the presence of multiple atomic tasks, association of tasks with KCs, and the potential for re-attempts.

3.3 Data collection of the case study

Generally, the process of obtaining knowledge from data consists of several steps, namely, data collection (student performance, engagement, behaviour, information about the learning environment, content and structure of the course), data cleaning including removal of any errors or inconsistencies, data analysis using statistical or machine learning techniques and finally knowledge discovery which can be used to make informed decisions and improve learning outcomes. In games, data analysis step could also include the usage of knowledge tracing algorithms which can be a powerful tool for improving learning outcomes by providing personalised support and feedback, and identifying areas of difficulty for targeted intervention. Next, we will explain the process of obtaining knowledge from the Shadowspect data.

Firstly, the data used for this work were collected as a part of an assessment machinery development. The MIT team recruited seven teachers in order to use the Shadowspect game for two hours in their 7th-grade and 10th-grade math and geometry classes. 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). For this paper, we used the data from 322 different students. Since our study was conducted with minors, and in accordance with ethical considerations and privacy laws, we deliberately avoided collecting any personal data from the students involved. The only identifier we used was a nickname chosen by the students themselves, which offers no insight into their individual demographic characteristics. Besides, the personal attributes of the students, such as gender, were not directly relevant to the primary goal of this study. Moreover, the omission of personal data ensures that we avoid collecting unnecessary and potentially sensitive information from minors.

It is also important to note that our study was conducted in schools across the United States, with the assistance of mathematics teachers participating in a fellowship program dedicated to the introduction of GBA in their classrooms. Thus, while we cannot provide specific demographic information, we can assure that the study was conducted in a real-world educational setting with a diverse set of students representing typical classrooms.

3.4 Knowledge Components in Shadowspect

The KCs in the game environment Shadowspect are the skills needed to complete a puzzle successfully. A math teacher-consultant was hired to identify these common core standards from the match curriculum supported by the Council of Chief State School Officers of the United States. The math teacher-consultant, following common core state standards of geometry [1], reviewed puzzles presented in the Shadowspect game and defined four main KCs, and most of the puzzles have the representation of three (GMD.4, CO.5 and CO.6) of them.

After developing the coding process, the following four common core standards were identified:

- **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 Figure 1 assume that to solve them, the student must have the following KCs: GMD.4, CO.5 and CO.6. It implies that in the case of these two puzzles, the importance of each KC means to have the same proportional weight (33%). In order to select the KCs for each puzzle and avoid the KCs co-occurrence issue, again, a math teacher-consultant was involved. Accordingly, most puzzles (90%) reflect the same idea of requiring the same three KCs because they are highly correlated, 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 an algorithm taking into account the correct weights of each KC in a puzzle, in our previous work [47], we assumed that all the present KCs had the same weight meaning that we considered that we did not have KCs prioritisation. In contrast, in the current work, we focus on doing an additional experiment of assigning priorities to certain KCs. Thus, we will compare the results of these two algorithmic configurations. We will describe the current formulas and description of the selected knowledge inference algorithms according to the new algorithmic configuration in the respective Section 3.6.

3.5 Algorithms implementation

The entire methodology process to predict the students' performance and model their actual knowledge is represented in Figure 2. As commented earlier, firstly, students of 7th-grade and 10th-grade math and geometry classes played Shadowspect as a part of their curriculum. Accordingly, in the second step, the data from students for this study were

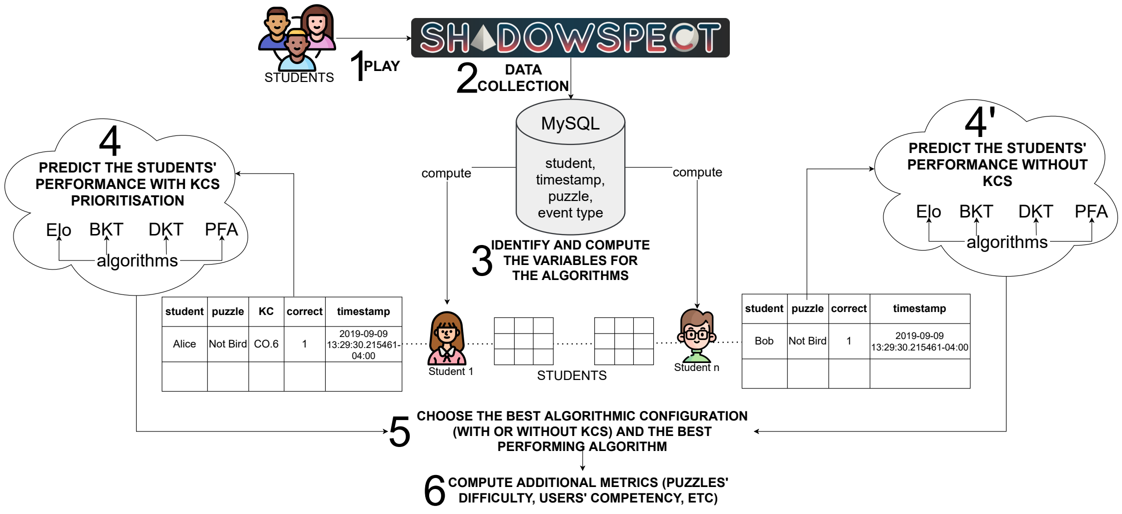

STUDENTS
1 PLAY
SHADOWSPECT
2 DATA COLLECTION
MySQL
student, timestamp, puzzle, event type
compute
compute
3 IDENTIFY AND COMPUTE THE VARIABLES FOR THE ALGORITHMS
4 PREDICT THE STUDENTS' PERFORMANCE WITH KCS PRIORITISATION
Elo BKT DKT PFA
algorithms
4' PREDICT THE STUDENTS' PERFORMANCE WITHOUT KCS
Elo BKT DKT PFA
algorithms
student puzzle KC correct timestamp
Alice Not Bird CO.6 1 2019-09-09 13:29:30.215461-04:00
Student 1
STUDENTS
Student n
student puzzle correct timestamp
Bob Not Bird 1 2019-09-09 13:29:30.215461-04:00
5 CHOOSE THE BEST ALGORITHMIC CONFIGURATION (WITH OR WITHOUT KCS) AND THE BEST PERFORMING ALGORITHM
6 COMPUTE ADDITIONAL METRICS (PUZZLES' DIFFICULTY, USERS' COMPETENCY, ETC)

Fig. 2. Overview of the methodology to predict the students' performance and evaluate the models

collected, and all interactions with the game were stored in a MySQL database. To be consistent with the common steps, we made the same data-related assumptions for all the algorithms. Respectively, we iterated through the input experiment database, 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 sorted given their respective timestamps, and the events of different students are interspersed, we made the separation between the events of each student. Then, in the third step, we computed if the student was correct or not in their 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 more than one submission of the 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 fourth step, we implemented the BKT, PFA, Elo and DKT algorithms for both algorithmic configurations (with and without KCs) in order to predict the students' performance based on the current modelling of each student. Next, the fifth step consisted in choosing the best-performing algorithm and the according algorithmic configuration (with or without KCs), and for it, the final sixth step was to compute additional metrics such as difficulty per puzzle or competency per student. With these values 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 help support both students' evaluation and potential adaptive learning approaches.

3.6 Algorithms

In this section, we describe the key approaches for knowledge inference, namely: **BKT**, **PFA**, **Elo** and **DKT** . We chose these algorithms because they represent the most common ones showing the best results applied to similar problems.

3.6.1 Bayesian Knowledge Tracing (BKT). BKT [9] 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 [12]. In this algorithm, only the first attempt on each task matters and learning is modelled by a discrete transition from an unknown to a known state. A fundamental assumption is that once a student knows a skill, they do not forget it.

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

- $p(L_0)$ - the probability that the student has prior knowledge meaning that they know 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.

These parameters can be fit from prior student data for each KC. Based on them, the inference is made about the student's probability of knowledge at time opportunity n, $p(L_n)$ which provides a powerful ability to track individual differences with each KC. The parameters and inferred probability of knowledge can also be used to predict the correctness of a student response with the following equations:

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

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

In the first place, an initial probability that the student has the knowledge is established because not all students have the same base, and this can have a negative influence if equality in previous knowledge is considered. Once the student was subjected to the task, a new probability appears, since now the knowledge is modified when exposed to the problem in question. We also consider the probabilities that the student will guess without having the skill acquired or that the student will not be correct actually having the knowledge. Notwithstanding, the puzzles in Shadowspect are less suitable for this type of algorithm because the guessing factor is reduced to almost 0. This algorithm presents an acceptable option for the problem we want to address. This is because it is a simple form of modelling the behaviour of each student and observe their competencies.

3.6.2 Performance Factors Analysis (PFA). The goal of the PFA model is to measure how much skill a student has during the learning process. It has considerable power to fit data and provides the adaptive flexibility to create the model overlay to be used adaptively by a tutor [36]. In the standard PFA model, the data about learner performance are used to compute a skill estimate. 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_{ij} - the prior successes for the KC of the student;
- f_{ij} - 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. Accordingly, it can be easily adapted to both algorithmic configurations – with KCs prioritisation and without. Moreover, the parameters in the PFA algorithm combined information from correctness with improvement from practice improvement, and we adjusted them for our particular case study.

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

$$m(i, j \in KCs, k \in Items, s, f) = \beta + \sum_{j \in KCs} (\gamma_j s_{ij} + \rho_j f_{ij}) \tag{4}$$

3.6.3 Elo rating system. Elo [38] is a skill calculation system that has been predominantly used to rank players, for example, in chess tournaments. It was developed for the purpose of measuring players' strength, but it was also applied in the context of educational research and was used for measuring both learner ability and task difficulty [34]. 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). In our case, we face a student with a task, but students never compete against each other. Accordingly, the expected outcome of the task and the probability of success for a user attempting a task are calculated. Based on them, we obtain the new Elo rating score [13].

First, we must obtain the probability that a student answers correctly a question by using a logistic function with both the competency 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 $correct_{si} \in \{0, 1\}$:

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

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). For our particular case study, $K = 0.05$. The following equations represent updates for both the competency of the student and the difficulty of the puzzle:

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

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

This Elo variant based on learning is a suitable algorithm given the characteristics that we have. The idea is to replace the pair of players that face each other and that we must analyse in the traditional algorithm for each student-question pair, simulating a competition in which the student wins if they answer the question and vice-versa.

3.6.4 Deep Knowledge Tracing (DKT). DKT [40] predicts performance on future items within a system based on Recurrent Neural Networks (RNNs) that are 'deep' in time to the task of knowledge tracing. Similar to BKT, this algorithm observes knowledge at the KC level and the correctness of each problem [56]. DKT uses the information from previous timestamps in order to make better future performance predictions while the student progresses through the task. Specifically, the algorithm uses Long Short-Term Memory to represent the latent knowledge space of students dynamically, determining how much information to remember from previous timestamps and how to combine that memory with information from the current timestamp. DKT uses large numbers of artificial neurons to represent latent knowledge state so the model can learn the latent knowledge state from data.

Traditional RNNs map an input sequence of vectors $x_1, \dots, x_T$, to an output sequence of vectors $y_1, \dots, y_T$. This is achieved by computing a sequence of 'hidden' states $h_1, \dots, h_T$ – successive encodings of relevant information from past observations that will be useful for future predictions. The variables are related using a simple network defined by the following equations:

$$h_t = tanh(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \tag{8}$$

$$y_t = sigmoid(W_{yh}h_t + b_y) \tag{9}$$

In DKT, both *tanh* and *sigmoid* functions are applied to each dimension of the input and parameterised by an input weight matrix W_{hx}, recurrent weight matrix W_{hh}, initial state h_0, and readout weight matrix W_{yh}. Biases for latent and readout units are represented by b_h and b_y.

3.7 Algorithms comparison applied to the geometry game environment

This section presents a methodological comparison of the above-described algorithms in order to analyse how they could perform better in our particular GBA case study, highlighting their advantages and disadvantages. Based on it, we will be able to select features for this context among algorithms.

Firstly, BKT is the most potent technique for cases when each question/task is primarily associated with a single KC. Therefore, observing students' performance on a given task only updates the probability of a single KC being in the learned state. Also, only the first attempt of each item is taken, so the algorithm throws out some information but, on the other hand, uses the most precise data because traditionally, the first answer that the student gives represents their real knowledge. Thus, this model fits well our case study, where we consider that puzzles do not have KCs prioritisation. However, one weak point of BKT is shown by Corbett and Bhatnagar [10], who stated that the knowledge tracing process tends to consistently overestimate students' performance by an average of 8%. The authors explained it by the following reasons: 1) it was hard for students to meet the tutor's expectations because they forgot the material they knew before, or 2) students were acquiring the knowledge that was not sufficient to perform adequately in a test. However, the advantage of this model is that it is easy to interpret the parameters as well as their effects on the models' performance.

In different circumstances, if multiple KCs at the same time are relevant to perform tasks, the rest of the algorithms are the most effective. As said before, unlike in BKT, in PFA, each item may involve multiple latent skills or KCs. Moreover, in PFA, each skill has a success learning rate and a failure learning rate, while in BKT, the learning rate is the same: success or failure. This is important in GBA and, in particular, in the considered puzzle game because students can learn a lot even while failing to solve the puzzle completely. In addition, PFA has an essential advantage over BKT because it does not consider errors in the decisive exercises, and it implies a more gradual modification. However, a disadvantage is that it ignores the order of the student's activities, leading to some information loss. Moreover, in the original formulation of the model, it assumes that the context in which a KC is learned has no effect on whether the concept is learned or not. However, we find an important strength to be considered for PFA in our environment: it softens the impact of incorrect answers. Therefore, the created model is more realistic and does not modify as much in the face of error. 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).

Another algorithm that allows associating each puzzle with several KCs is the Elo algorithm. It is easy to implement in educational systems and can be easily used in an online setting. Also, Elo is not restricted to binary results and can be used for problems where time is the only performance measure which is a common situation in GBA. Besides, the implementation and the adaptation of the Elo algorithm to the data are not that complicated. The algorithm has few adjustment parameters, and it is also computationally very simple and fast [49].

Finally, the DKT algorithm proved to show decent results in comparison with other approaches. While the key reason for the success of this model is its ability to capture the sequential dependencies among questions embedded in the question-answer sequences [51], it suffers from two major problems. Firstly, the model fails to reconstruct the observed input. As a result, even when a student performs well on a KC, the prediction of that KC's mastery level could decrease instead, and vice versa. Secondly, the predicted performance for KCs across time-steps is not consistent. This is undesirable and unreasonable because student's performance is expected to evolve gradually over time [55].

Considering all advantages and disadvantages of the selected algorithms, we decided to adapt and apply all of them to our case study though their implementation and results interpretation will heavily rely on the stated drawbacks.

3.8 Classifier Metrics

We will be using each algorithm to obtain a standardised numerical value between 0 and 1 representing the geometry capabilities of every student according to each KC based on the history of each student's interactions with the activity. After exploring the work performed by Pelánek [37], who carried out 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 to order data points 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 considering 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 in the configuration without KCs and then, after prioritising certain KCs. We will compare the performance of selected algorithms by the use of selected metrics. Finally, we will choose the best performing algorithm and conduct additional experiments in order to present a more detailed overview.

4.1 Algorithms metrics comparison (RG1)

w/o KCs	**AUC**	**Accuracy**	**F1 Score**
BKT	0.78	0.87	0.93
PFA	0.78	0.87	0.93
Elo	0.89	0.94	0.97
DKT	0.79	0.88	0.93

with KCs	**AUC**	**Accuracy**	**F1 Score**
BKT	0.72	0.84	0.91
PFA	0.72	0.85	0.92
Elo	0.87	0.94	0.97
DKT	0.74	0.85	0.91

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

From Table 1, we can observe the fact that all the algorithms show decent results for the selected metrics, in both cases: before and after declaring the KCs. In the case without KCs, the accuracy of the BKT and PFA algorithms is identical, and DKT shows a slightly better result while the Elo algorithm outperforms them by ~7-8%. We also observe a similar pattern in the AUC metric; however, in this case, Elo performs better by a more significant value of ~10-11%. 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 take into account any imbalance when looking at the accuracy and the AUC metrics. Finally, since we have a skewed sample distribution, in Section 3.8, 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 (while Elo again outperforms by ~4%). Therefore, we can deduce that the BKT, PFA, Elo, and DKT algorithms are precise and robust.

After prioritising KCs, we still see the acceptable metrics outcomes of all the algorithms. However, all of them narrowly worsened, and the metric that was affected the most is AUC. On the other hand, accuracy and F1 score decreased only by ~1-3%.

To sum up, the results reveal that all the algorithms show reasonable outcomes considering our case study on a game environment, and therefore we believe that our adaption into Shadowspect was successful. While BKT, PFA and DKT perform sufficiently well, we can observe that the Elo algorithm outperforms them in the most critical metrics – AUC and F1 score. Therefore, we conclude that for both cases (with and without KCs), the most predictive model is Elo, which outperforms the overall accuracy of BKT, PFA and DKT by ~8% and F1 score by ~5% without KCs. Simultaneously, it shows pretty similar results after KCs prioritisation decreasing the AUC metric only by ~2% while other algorithms worsened their metrics results much more.

One of the primary contributions of our work is adapting knowledge inference algorithms to the context of the Shadowspect game. These algorithms are often used in ITSs but our adaptation to games is novel. Therefore, there are not that many studies published with which we can perform a fair comparison. Specifically, recently, there was not performed any similar study including the Elo algorithm. However, other studies in different settings have similar performance and our results coincide with other works in the fact that generally DKT outperforms more traditional models like BKT and PFA [45].

4.2 Elo algorithm detailed results (RG2)

In this section, we will present the additional results of the best-performing algorithm. As we concluded earlier, Elo is showing the best results, and since the algorithm changes the student parameter and an item parameter every time it encounters a new experience taking the difference between the students' actual response and predicted response and weighting it by a coefficient (see Equations 6 and 7), we see potential in performing additional experiments. Therefore, we will discuss the Elo algorithm in detail, presenting additional metrics such as accuracy and difficulty computed for each puzzle and accuracy and average active time obtained for each student.

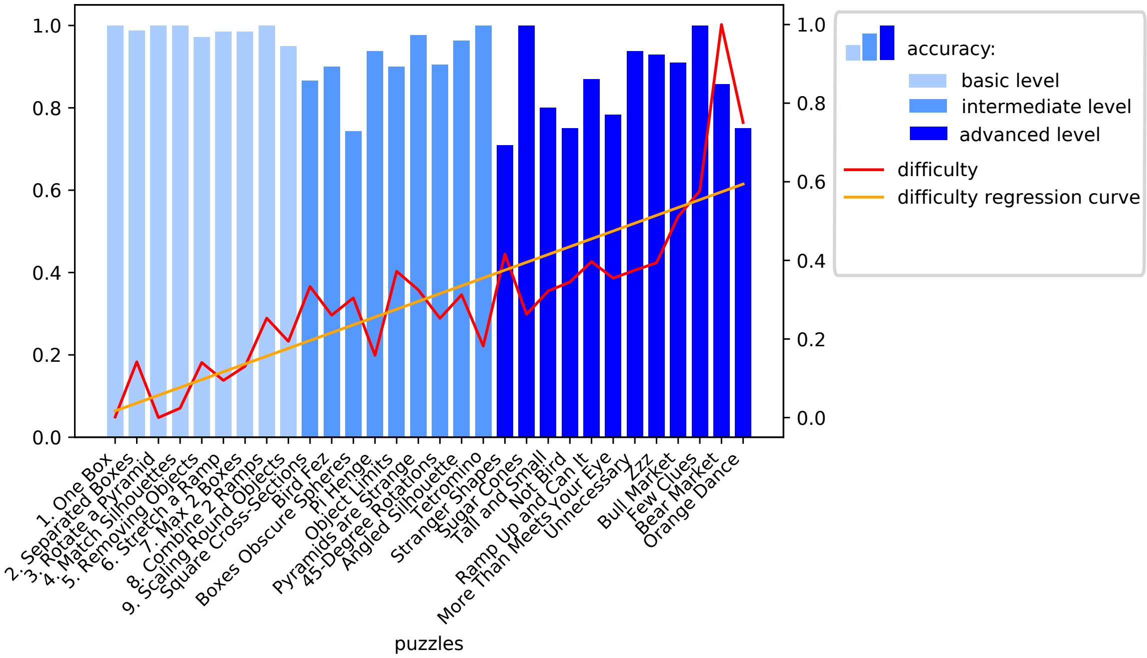


Fig. 3. Elo accuracy and difficulty results per puzzle split by Shadowspect designers' difficulty

In Figure 3, we represent the accuracy metric per puzzle, the puzzle difficulty generated by the Elo algorithm (see Equation 7) and the difficulty regression curve, additionally splitting the puzzles by Shadowspect designers' difficulty (basic, intermediate and advanced). Accordingly, the x-axis is ordered based on the sequence order of puzzles presented in the Shadowspect game, the left y-axis represents the accuracy, and the right y-axis stands for both the difficulty and the difficulty regression curve. We can observe the fact that the curve of the difficulty goes up at the same time as the difficulty levels. This is a feature implemented in a lot of games in order to maintain the user engaged and challenged as the game advances. This is called intrinsic motivation while studying directed towards the enjoyment of experience [53].

We can notice that it is easier to predict the student's performance in the most manageable basic levels, while in more sophisticated levels, the accuracy decreases due to a higher variance in students' capabilities to solve each puzzle. It can be readily explained by the fact that in Shadowspect, the first nine puzzles are used for teaching the fundamentals of the game, while later on, the users are making their own decisions to solve the puzzles. We can observe that the difficulty of several puzzles (e.g., Bear Market and Orange Dance) stands out above the rest. Accordingly, it is harder to predict the students' performance in them.

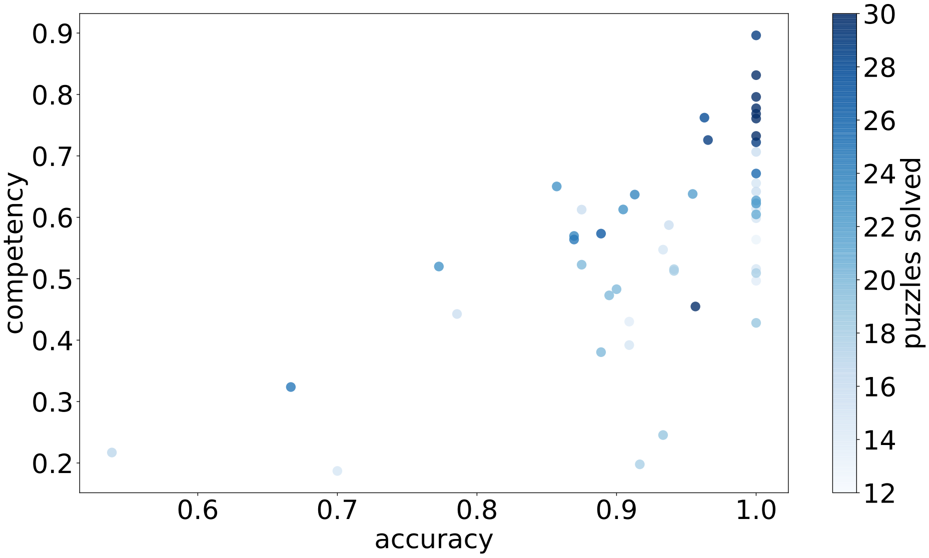


Fig. 4. Elo average activity time and accuracy results per user

In Figure 4, we represent calculations of the accuracy metric per user and their competency, specifying how many puzzles each student solved. Accordingly, the x-axis represents the accuracy while the y-axis stands for the competency. The intensity of the dot's colour codifies the number of puzzles the student solved (the darker the dot is, the more puzzles the student solved). It is essential to say that we selected only those students who spent a reasonable amount of time playing the game (more than 60 seconds of active time per puzzle). In total, 85 students met this criterion. At the same time, we only analysed the behaviour of those students who attempted to solve at least 12 puzzles because we suppose that students solve the puzzles by difficulty starting with an intermediate level which teaches the game functionality and finishing with an advanced level for the students who already have enough experience. As stated before, the intermediate level includes a total of nine puzzles, and we consider that it is reasonable to include in the study only those students who started to take their personal decisions in the game starting from the tenth puzzle. In total, 51 students met this criterion, and the same amount of students met all the criteria stated before.

Moreover, following the Equation 6, we calculated the competency of each student for all KCs, which is obtained based on the history of each student's interactions with the activity. With this value in mind, the teacher knows the learner's ability before conducting a formal assessment. Thus, it is intended to act both individually and collectively, e.g., if strange behaviours appear at an individual level, the teacher can intervene to see the cause and help the student improve. Through these measures, we seek to collaborate both in the evaluation of students and in adaptive learning.

5 DISCUSSION

In this section, we will present a discussion following the obtained results. Firstly, we will generally discuss our findings. Next, we will raise two critical topics: the importance of KCs and the feasibility of using the Elo algorithm for formative assessment. Finally, we will raise the limitations of our work.

5.1 Performance of the algorithms

Based on the metrics comparison presented in Table 1, we are assured that our results proved that it was possible to effectively adapt the BKT, PFA, Elo and DKT algorithms to a GBA. Therefore, our study conducted on Shadowspect can serve for future research conducted on games with similar characteristics. It demonstrates a scalable approach to applying knowledge inference algorithms in diverse game-based learning environments, extending beyond the specific context of geometry learning. We did not find any work adapting and/or applying all mentioned above knowledge inference algorithms to one case study and particularly to GBA. Therefore, we do not have any reference with which to compare our outcomes. However, in Section 2.2 we mentioned several works comparing other algorithms. For example, Gervet et al. [17] stated that DKT could perform better when precise temporal information matters most. In our case study, DKT also showed decent results, but Elo outperformed it. We explain it with the fact that this was the algorithm customised the most to GBA. On the other hand, we found a unique work whose authors applied and compared several knowledge algorithms (BKT and DBNs) to GBA [11]. They concluded that both models are valuable and informative, while in our case, BKT was not the algorithm showing the best results in modelling the real knowledge of students. This could be expected because the game presented in their work has different functionality and goals.

Considering another goal of the paper at hand of comparing two algorithmic configurations – with KCs prioritisation and without – we are not aware of any study exploring this issue, and therefore, we draw our conclusions in the following subsection.

5.2 Do KCs matter?

In our work, we compared the performance of the BKT, PFA, Elo and DKT algorithms in the context of the geometry game environment Shadowspect. Additionally, considering our particular case study, we decided to do complementary experiments on separating KCs and prioritising them for certain puzzles. After comparing the results (see Table 1), we observed the fact that the algorithms showed better results before the separation of KCs. This could be explained by the fact that assigning priorities to KCs, in fact, signifies information loss. To illustrate, initially, the puzzle "Bull Market" encompassed all four KCs. However, once the expert determined that the prioritised KC for this puzzle was MG.1, we effectively lost the representation of the other three KCs, which constitutes 75% of the original information. However, our output does not signify that every case study would not benefit from KCs separation.

Moreover, we concluded that Elo is the best performing algorithm. We saw that its metrics did not deteriorate significantly after we prioritised some KCs (accuracy and F1 score remained the same while AUC decreased from 0.89 to 0.87). This means that each particular case is different, so the choice of according metrics and the usage of KCs will vary. The geometry game Shadowspect was designed explicitly for developing geometry capabilities where KCs are highly correlated with one another. We are of the opinion that there could be an effect on separating KCs in other studies. ITS environments could significantly benefit, as their goal is to provide immediate and customised feedback to students. This is also explained by the reason that ITSs are constrained environments while games are much more open environments without a clear balance between reliability and entertainment. Moreover, it could be helpful if the feedback provided to learners could be separated by focusing on every particular KC. In this way, both teachers and students would know what exact skills they need to polish. Nevertheless, the generalizability of our results should be approached carefully since the outcomes we obtained do not guarantee similar results in all studies adapting knowledge inference algorithms to GBA.

5.3 Use of the knowledge inference algorithms for formative assessment and GBA within the example of Elo

Following the results presented in Section 4, we are confident that the Elo algorithm can be used for formative assessment. There are two main simplest scenarios that could benefit from it. The first one was described by Pankiewicz and Bator [34], who compared methods of task difficulty estimations suitable for use within online learning environments for the purpose of adaptive item sequencing. Accordingly, in two groups of learners (beginners and experienced learners) Elo outperformed two other methods, namely, proportion correct and learner feedback. The second scenario is the estimation of the learner's proficiency level [52].

As we saw in Figure 4, there is some variance in students' performance throughout the puzzles. More specifically, there are some students whose results significantly negatively differ from others. Among the various student cases, we can distinguish between top performers and those who are struggling. It is easier to predict the behaviour and outcomes of students solving the puzzles better than the average. Additionally, it is not a trivial task to predict the performance of those students whose competence is low. We can also see that the competency of students increases while they solve more puzzles. When a teacher analyses the performance of a particular student and encounters such a case, they must investigate the cause. Interpreting these cases can be tricky, as various factors might be at play: (1) the student may have behaved in a thoughtless manner, causing the data to stand out; (2) the student might be struggling with a particular aspect that appears in that puzzle, indicating an individual-level problem; or (3) the specific puzzle may present a greater level of difficulty compared to the others (see Figure 3).

In this sense, there are several challenges of implementing the Elo algorithm in real scenarios. These challenges stem from the intricacies of the formative assessment process, an example of which is [41]. The authors developed a visualisation dashboard that can play the role of a tool supporting teachers using educational games in the classroom. In this case, teachers are able to receive useful information derived from the knowledge inference algorithms via the dashboard and then provide personalised feedback to students who accordingly can improve their skills based on the previous experience. Respectively, the first challenge is to establish clear instructions for the teachers especially considering already existing obstacles while handling big classes. For example, if it is evident that there is one puzzle in which most of the students fail or have difficulties, the instructor can explain how this particular task is solved. However, if distinct students have doubts about different puzzles, the situation is more unclear. On the other hand, it is compensated with the fact that the algorithm can be applied in real-time scenarios because the competency of students and the difficulty of puzzles are evaluated for each particular case.

Also, with the metrics presented in 4.2, we can address two critical problems – academic dishonesty (or cheating) and teachers' trustworthiness. Cheating is a common issue in the classroom being yet unresolved [3]. The most spread reason for it is to gain a competitive advantage in the race of position The most common reason for cheating is to gain a competitive advantage in the quest for status or power [50]. In regular classrooms, students cheat in order to achieve better results and, therefore, higher grades. In this case, it is important to early identify these students and understand the reasons. More than that, to avoid these situations, it is vital that teachers do not teach the test (similar to the exam questions) but introduce the thinking strategies and do not neglect any topic that will not be tested on. Also, another challenge brings the existence of the phenomena called “gaming the system", which represents the idea that a student attempts to succeed in an educational environment by exploiting properties of the help and feedback tools of the system rather than by learning the material [4]. We saw these situations in Shadowspect when students, while solving the puzzles, were explicitly asking for help by repeatedly clicking on the submit button in order to receive feedback on the solution.

It is also important to mention that there are various types of students. For example, shallow learners – they do well on tests they learned from the system but are not able to transfer their knowledge to new contexts [5]. Also, there are students that just randomly insert solutions with erratic behaviours. This could be happening due to several reasons, including lack of educational self-drive, frustration during the learning process or simply that the student dislikes the subject matter. Accordingly, it is harder to predict the behaviour and model the real knowledge of such learners. This brings the limitations of our work presented in the next section.

In conclusion, the ability to predict students' performance with knowledge tracing algorithms is of global interest because it has the potential to revolutionise the field of education and make learning more effective, efficient, and personalised. This is supported by the fact that these algorithms can improve learning outcomes, lead to a more efficient resource allocation and offer cost-effective solutions [31]. Additionally, the development of knowledge tracing algorithms requires interdisciplinary collaboration and innovation in the fields of computer science, education, psychology, and data analytics which can lead to new discoveries, approaches, and technologies that can benefit a wide range of industries and fields [54]. Finally, education is a critical tool for addressing global challenges such as poverty, inequality, and sustainable development. Knowledge tracing algorithms can help make education more effective, accessible, and inclusive, thereby contributing to the achievement of these goals.

5.4 Educational implications

Given that the knowledge tracing algorithms are performing effectively, it follows that we are able to accurately evaluate one's competence in geometry and spatial reasoning. This outcome is significant not only due to the importance of these skills, but also because of their correlation with other advanced cognitive abilities and problem-solving aptitudes [7]. Specifically, individuals who demonstrate proficiency in spatial reasoning often exhibit similar aptitudes in solving a diverse range of problems [21]. Therefore, this assessment machinery can be helpful to detect and provide feedback on problem-solving skills that are key for the 21st century society. In this way, there are several practical and educational implications of using knowledge tracing algorithms in games to predict student performance and model their real knowledge:

- **Personalised Learning**. Knowledge tracing algorithms are able to help in the task of personalising the learning experience for each student. By modelling each student's knowledge level and predicting their performance, the game can be adapted to meet their individual learning needs. In result, we can obtain increased engagement and motivation, and ultimately students can achieve better learning outcomes.
- **Early Intervention**. Knowledge tracing algorithms can identify areas of difficulty or misconceptions early on, allowing for targeted intervention and support before the problem becomes more serious. Accordingly, this can be beneficial in preventing students from falling behind and ensure that they have a solid foundation of knowledge before moving on to more advanced concepts.

- **Adaptive Assessment**. Knowledge tracing algorithms can also adapt the assessment process to each student's level of knowledge and skill based on the prediction of each student's performance.
- **Improved Instructional Design**. The data generated by knowledge tracing algorithms can be used to inform instructional design and improve the overall effectiveness of the games. By analysing patterns and trends in student performance and behaviour, designers can make informed decisions about how to improve the game mechanics, learning objectives, and instructional strategies.

In summary, the use of knowledge tracing algorithms in games has the potential to significantly improve the learning experience for students, while also providing valuable insights to instructors and designers [25]. These algorithms can help to personalise the learning experience, identify areas of difficulty, and inform instructional design since they are able to model the knowledge of every student and predict their performance. In this way, the ability to predict students' performance with knowledge tracing algorithms can have a significant impact on the design and use of games in several ways. Apart from adapting the content, difficulty level, and feedback to the individual needs and learning progress of each student and accordingly, providing real-time feedback to students on their performance and progress, our study suggests two additional recommendations. Firstly, knowledge tracing algorithms can be used to assess students' knowledge and skills in a more accurate and objective way than traditional assessments. Also, they can be used to optimise the design of games by identifying the most effective learning strategies, game mechanics, and content for different types of learners. This can help game designers and developers create more engaging and effective learning experiences, as well as improve the overall quality of games.

Beyond these practical implications, our study also contributes to the broader field of educational research in several significant ways. Firstly, our research exemplifies how theoretical algorithms from the field of data analytics and knowledge tracing can be practically applied to real-world educational contexts. Secondly, by using these algorithms, educators gain a more complete understanding of each student's learning path which could help to target the resources more effectively and may aid in lesson planning and resource allocation. Thirdly, the approach suggested in our research could lead to more inclusive education by ensuring that each student's unique learning needs are addressed. Lastly, our findings have implications beyond the classroom and can inform educational policy and curriculum design. By demonstrating the effectiveness of gamified learning and knowledge tracing, our study can contribute to discussions on modernising educational approaches and curricula. In conclusion, our study not only introduces a novel method for applying knowledge tracing algorithms to geometry-based games, but also offers valuable insights and implications for educators, game designers, policymakers, and researchers in the field of education.

5.5 Limitations

By exploring the example of the Elo algorithm applied to the Shadowspect case study, we conclude that other knowledge inference algorithms could also be adapted to formative assessment and GBA. However, following the challenges of implementing them in formative assessment and GBA mentioned in the previous section, we will mention the limitations of our work.

First of all, our particular case study shows the results of the experiments performed by using data collected from 322 different students. We are confident that this represents a decent sample size in order to conclude reliable results. However, these results come from the sample from a single context since these data were collected as a part of an assessment machinery development. Respectively, in order to have more diverse data, it is needed to collect data from users that were not solving the puzzles as a part of their curriculum. Therefore, we are not confident that our algorithms adaptation and outcomes would generalise to other contexts.

Moreover, at the beginning of our work, we questioned if the results of our experiments would be different while comparing two algorithmic configurations – with KCs prioritisation and without. It is worth mentioning that the geometry game Shadowspect requires the mastery of four KCs that are highly correlated to each other, and moreover the data we tagged had overlapping between KCs. Accordingly, the limitation that we faced consisted in the fact that the puzzles were designed before taking into account the KCs. Respectively, we are of the opinion that it could be beneficial to consider KCs from the very beginning of designing game levels. In this way, our study can serve as a base only for similar GBAs where the KCs are focused on similar skills.

6 CONCLUSIONS AND FUTURE WORK

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 with learner perspectives. This work presents a novel adaptation analysis and comparison of four knowledge inference algorithms – namely, BKT, PFA, Elo and DKT – applied to a geometry game environment to model learners' latent knowledge. We analysed the results of two algorithmic configurations – one with prioritisation of KCs, the skills needed to successfully complete a puzzle, and one without this prioritisation. We measured the performance of the algorithms mentioned above by examining the following metrics: AUC, accuracy and F1 score. We observed that all the algorithms showed more decent results before we separated the KCs, which could be explained by the fact that assigning priorities to KCs esentially signifies information loss. Moreover, Shadowspect was designed explicitly for developing geometry capabilities where KCs are highly correlated. Among all the algorithms, we found Elo to have the best predictive power and the ability to model the real knowledge of students. 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 real knowledge of students. In turn, this confirms that all the four analysed algorithms are suitable for application in formal education to improve teaching, learning, and organisational efficiency.

Besides, we are confident that this work can motivate teachers and students to use game-design elements for the learning process as discussed in Section 5.4. This is due to the fact that the paper at hand proved that it is possible not only to make the educational process enjoyable through playing but also to meet the desires and needs of every student. We discussed the possible scenarios of the Elo algorithm application in formative assessment and identified that it could be used for the estimations of task difficulty and the learner's proficiency. Moreover, this experience could also be transferred into not formal educational settings with new innovative products.

As far as we know, this is the first research conducted on adapting and applying and comparing the above-mentioned knowledge inference models in GBA to predict learners' performance. Besides, there are several possible extensions to this research. For example, further experiments are needed to conclusively determine whether KCs prioritisation might not work well in GBA. Moreover, Pardos et al. [35] concluded that combining multiple approaches through ensemble selection can be more effective for large data sets than single models and accordingly, this approach could be explored further for every particular case study.

ACKNOWLEDGMENTS

We thank Pedro A. Martínez for his help with the implementation of the Elo algorithm.

This work was partially supported by REASSESS project (grant 21948/JLI/22), funded by the Call for Projects to Generate New Scientific Leadership, included in the Regional Program for the Promotion of Scientific and Technical Excellence Research (2022 Action Plan) of the Seneca Foundation, Science and Technology Agency of the Region of Murcia.

REFERENCES

[1] 2022. Common Core State Standards of Geometry. http://www.corestandards.org/Math/Content/G/ . The access date: 01.08.2022.

[2] Ali Alkhatlan and Jugal Kalita. 2018. Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments. *arXiv preprint arXiv:1812.09628* (2018). https://doi.org/10.48550/ARXIV.1812.09628

[3] Andi Asrifan, Abd Ghofur, and Nur Azizah. 2020. Cheating behavior in EFL classroom (a case study at elementary school in Sidenreng Rappang Regency). *OKARA: Jurnal Bahasa dan Sastra* 14, 2 (2020), 279–297. https://doi.org/10.19105/ojbs.v14i2.4009

[4] Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research* 19, 2 (2008), 185–224.

[5] Ryan SJD Baker, Sujith M Gowda, Albert T Corbett, and Jaclyn Ocumpaugh. 2012. Towards Automatically Detecting Whether Student Learning Is Shallow. In *International Conference on Intelligent Tutoring Systems*. Springer, Springer Berlin Heidelberg, 444–453.

[6] Stacey Brull and Susan Finlayson. 2016. Importance of gamification in increasing learning. *The Journal of Continuing Education in Nursing* 47, 8 (2016), 372–375. https://doi.org/10.3928/00220124-20160715-09

[7] Jeffrey Buckley, Niall Seery, and Donal Canty. 2018. Investigating the use of spatial reasoning strategies in geometric problem solving. *International Journal of Technology and Design Education* 29, 2 (March 2018), 341–362. https://doi.org/10.1007/s10798-018-9446-3

[8] 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. https://doi.org/10.1007/s40593-020-00202-6

[9] 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. https://doi.org/10.1007/BF01099821

[10] Albert T Corbett and Akshat Bhatnagar. 1997. Student Modeling in the ACT Programming Tutor: Adjusting a Procedural Learning Model With Declarative Knowledge. In *User Modeling*. Springer, Springer Vienna, 243–254.

[11] Yang Cui, Man-Wai Chu, and Fu Chen. 2019. Analyzing Student Process Data in Game-Based Assessments with Bayesian Knowledge Tracing and Dynamic Bayesian Networks. *Journal of Educational Data Mining* 11, 1 (2019), 80–100.

[12] 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, Springer Berlin Heidelberg, 406–415.

[13] Ole Halvor Dahl and Olav Fykse. 2018. *Combining Elo Rating and Collaborative Filtering to improve Learner Ability Estimation in an e-learning Context*. Master's thesis. NTNU.

[14] Dalila Durães, Rámon Toala, Filipe Gonçalves, and Paulo Novais. 2019. Intelligent tutoring system to improve learning outcomes. *AI Communications* 32, 3 (2019), 161–174. https://doi.org/10.3233/AIC-190624

[15] Sandeep Dutta and Eric Gros. 2018. Evaluation of the impact of deep learning architectural components selection and dataset size on a medical imaging task. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, Jianguo Zhang and Po-Hao Chen (Eds.), Vol. 10579. International Society for Optics and Photonics, SPIE, 240 – 253. https://doi.org/10.1117/12.2293395

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

[17] Theophile Gervet, Ken Koedinger, Jeff Schneider, and Tom Mitchell. 2020. When is Deep Learning the Best Approach to Knowledge Tracing? *JEDM| Journal of Educational Data Mining* 12, 3 (2020), 31–54. https://doi.org/10.5281/zenodo.4143614

[18] 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. https://doi.org/10.1016/j.jecp.2014.01.002

[19] Manuel J. Gomez, José A. Ruipérez-Valiente, and Félix J. García Clemente. 2022. A Systematic Literature Review of Game-based Assessment Studies: Trends and Challenges. *IEEE Transactions on Learning Technologies* (2022), 1–16. https://doi.org/10.1109/TLT.2022.3226661

[20] Jose Gonzalez-Brenes, Yun Huang, and Peter Brusilovsky. 2013. Fast: Feature-Aware Student Knowledge Tracing. In *Proceedings of NIPS 2013 Workshop on Data Driven Education*. University of Pittsburgh. http://d-scholarship.pitt.edu/20353/

[21] Zachary Hawes and Daniel Ansari. 2020. What explains the relationship between spatial and mathematical skills? A review of evidence from brain and behavior. *Psychonomic Bulletin Review* 27, 3 (Jan. 2020), 465–482. https://doi.org/10.3758/s13423-019-01694-7

[22] Yoon Jeon Kim and Dirk Ifenthaler. 2019. Game-Based Assessment: The Past Ten Years and Moving Forward. In *Game-Based Assessment Revisited*. Springer International Publishing, 3–11. https://doi.org/10.1007/978-3-030-15569-8_1

[23] 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. https://doi.org/10.1111/j.1551-6709.2012.01245.x

[24] Tamara G Kolda and Brett W Bader. 2009. Tensor Decompositions and Applications. *SIAM Rev.* 51, 3 (2009), 455–500. https://doi.org/10.1137/07070111X

[25] Jeanine Krath, Linda Schürmann, and Harald F.O. von Korflesch. 2021. Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior* 125 (2021), 1–33. https://doi.org/10.1016/j.chb.2021.106963

[26] Kandamaran Krishnamurthy, Nikil Selvaraj, Palak Gupta, Benitta Cyriac, Puvin Dhurairaj, Adnan Abdullah, Ambigga Krishnapillai, Halyna Lugova, Mainul Haque, Sophie Xie, and Eng-Tat Ang. 2022. Benefits of gamification in medical education. *Clinical Anatomy* 35, 6 (2022), 795–807. https://doi.org/10.1002/ca.23916

[27] 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. https://doi.org/10.1007/s11423-017-9542-1

[28] Qi Liu, Shuanghong Shen, Zhenya Huang, Enhong Chen, and Yonghe Zheng. 2021. A Survey of Knowledge Tracing. https://doi.org/10.48550/ARXIV.2105.15106

[29] Ting Long, Yunfei Liu, Jian Shen, Weinan Zhang, and Yong Yu. 2021. Tracing Knowledge State with Individual Cognition and Acquisition Estimation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3404835.3462827

[30] Pedro A Martínez, Manuel J G Moratilla, José A Ruipérez-Valiente, Gregorio M Pérez, and YJ Kim. 2020. Visualizing Educational Game Data: A Case Study of Visualizations to Support Teachers. (Jun 2020). https://doi.org/10.35542/osf.io/9pz68

[31] Sein Minn, Jill-Jênn Vie, Koh Takeuchi, Hisashi Kashima, and Feida Zhu. 2022. Interpretable Knowledge Tracing: Simple and Efficient Student Modeling with Causal Relations. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 12810–12818. https://doi.org/10.1609/aaai.v36i11.21560

[32] Hafidi Mohamed and Mahnane Lamia. 2018. Implementing flipped classroom that used an intelligent tutoring system into learning process. *Computers & Education* 124 (2018), 62–76. https://doi.org/10.1016/j.compedu.2018.05.011

[33] Koki Nagatani, Qian Zhang, Masahiro Sato, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma. 2019. Augmenting Knowledge Tracing by Considering Forgetting Behavior. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3101–3107. https://doi.org/10.1145/3308558.3313565

[34] 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. https://doi.org/10.15219/em82.1444

[35] Zachary A Pardos, Sujith M Gowda, Ryan SJd Baker, and Neil T Heffernan. 2012. The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software. *ACM SIGKDD Explorations Newsletter* 13, 2 (2012), 37–44. https://doi.org/10.1145/2207243.2207249

[36] 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

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

[38] Radek Pelánek. 2016. Applications of the Elo rating system in adaptive educational systems. *Computers & Education* 98 (2016), 169–179. https://doi.org/10.1016/j.compedu.2016.03.017

[39] Radek Pelánek. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction* 27, 3 (2017), 313–350. https://doi.org/10.1007/s11257-017-9193-2

[40] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. 28 (2015). https://proceedings.neurips.cc/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf

[41] José A Ruipérez-Valiente, Manuel J Gomez, Pedro A Martínez, and Yoon Jeon Kim. 2021. Ideating and Developing a Visualization Dashboard to Support Teachers Using Educational Games in the Classroom. *IEEE Access* 9 (2021), 83467–83481. https://doi.org/10.1109/ACCESS.2021.3086703

[42] 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. https://doi.org/10.1016/j.compedu.2020.104008

[43] 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.

[44] Shaghayegh Sahebi, Yun Huang, and Peter Brusilovsky. 2014. Parameterized exercises in java programming: using knowledge structure for performance prediction. In *The second Workshop on AI-supported Education for Computer Science (AIEDCS)*. University of Pittsburgh, 61–70.

[45] Sami Sarsa, Juho Leinonen, and Hellas. Arto. 2022. Empirical Evaluation of Deep Learning Models for Knowledge Tracing: Of Hyperparameters and Metrics on Performance and Replicability. *Journal of Educational Data Mining* 14, 2 (Sept. 2022), 32–102. https://doi.org/10.5281/zenodo.7086179

[46] Richard Scruggs, Ryan S. Baker, and Bruce M. McLaren. 2019. Extending Deep Knowledge Tracing: Inferring Interpretable Knowledge and Predicting Post-System Performance. https://doi.org/10.48550/ARXIV.1910.12597 arXiv:1910.12597 [cs.CY]

[47] Sofia Strukova, José A. Ruipérez-Valiente, and Félix Gómez Mármol. 2021. Data-Driven Performance Prediction in a Geometry Game Environment. In *Proceedings of the Conference on Information Technology for Social Good* (Roma, Italy) *(GoodIT '21)*. Association for Computing Machinery, New York, NY, USA, 283–288. https://doi.org/10.1145/3462203.3475905

[48] Sofia Strukova, José A. Ruipérez-Valiente, and Félix Gómez Mármol. 2022. A Survey on Data-Driven Evaluation of Competencies and Capabilities across Multimedia Environments. *International Journal of Interactive Multimedia and Artificial Intelligence* (2022). https://doi.org/10.9781/ijimai.2022.10.004

[49] 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 International Publishing, 307–323. https://doi.org/10.1007/978-3-030-18480-3_16

[50] Zoran Vojinovic and Michael B Abbott. 2012. *Flood risk and social justice*. IWA Publishing.

[51] Zhiwei Wang, Xiaoqin Feng, Jiliang Tang, Gale Yan Huang, and Zitao Liu. 2019. Deep Knowledge Tracing with Side Information. In *International Conference on Artificial Intelligence in Education*. Springer, Springer International Publishing, 303–308.

[52] Kelly Wauters, Piet Desmet, and Wim Van Noortgate. 2010. Monitoring learners' proficiency: weight adaptation in the Elo rating system. In *Educational Data Mining 2011*.

[53] Maria Meiha Wong and Mihaly Csikszentmihalyi. 1991. Motivation and Academic Achievement: The Effects of Personality Traits and the duality of Experience. *Journal of Personality* 59, 3 (1991), 539–574. https://doi.org/10.1111/j.1467-6494.1991.tb00259.x

[54] Sheng Xu, Manfang Sun, Weili Fang, Ke Chen, Hanbin Luo, and Patrick X.W. Zou. 2023. A Bayesian-based knowledge tracing model for improving safety training outcomes in construction: An adaptive learning framework. *Developments in the Built Environment* 13 (2023), 100111. https://doi.org/10.1016/j.dibe.2022.100111

[55] Chun-Kit Yeung and Dit-Yan Yeung. 2018. Addressing Two Problems in Deep Knowledge Tracing via Prediction-Consistent Regularization. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (London, United Kingdom) *(L@S '18)*. Association for Computing Machinery, New York, NY, USA, Article 5, 10 pages. https://doi.org/10.1145/3231644.3231647

[56] Liang Zhang, Xiaolu Xiong, Siyuan Zhao, Anthony Botelho, and Neil T Heffernan. 2017. Incorporating Rich Features into Deep Knowledge Tracing. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale* (Cambridge, Massachusetts, USA) *(L@S '17)*. Association for Computing Machinery, New York, NY, USA, 169–172. https://doi.org/10.1145/3051457.3053976