The Affordances of Multivariate Elo-based Learner Modeling in Game-based Assessment

José A. Ruipérez-Valiente, *Senior Member, IEEE,* Yoon Jeon Kim, Ryan S. Baker, Pedro A. Martínez, and Grace C. Lin

***Abstract*—Previous research and experiences have indicated the potential that games have in educational settings. One of the possible uses of games in education is as game-based assessments (GBA), using game tasks to generate evidence about skills and content knowledge that can be valuable. There are different approaches in the literature to implement the assessment machinery of these GBA, all of them having strengths and drawbacks. In this paper, we propose using multivariate Elo-based learner modeling, as we believe it has a strong potential in the context of GBA for three aims: 1) to simultaneously measure students competence across several knowledge components in a game, 2) to predict task performance, and 3) to estimate task difficulty within the game. To do so, we present our GBA Shadowspect, which is focused on solving geometry puzzles, and we depict our implementation using data collected from several high schools across the USA. We obtain high performing results (AUC of 0.87) and demonstrate that the model enables analysis of how each student's competency evolves after each puzzle attempt. Moreover, the model provides accurate estimations of each task's difficulty, enabling iterative improvement of the game design. This study highlights the potential that multivariate Elo-based learner modeling has within the context of GBA, sharing lessons learned, and encouraging future researchers in the field to consider this algorithm to build their assessment machinery.**

***Index Terms*—Game-based assessment, knowledge inference, learning analytics, educational technology, K-12 education.**

I. Introduction

GAMES in general have been depicted as a considerable asset for learning [1]–[3]; this includes both those games that have been explicitly designed for learning purposes by mapping the mechanics and contents of the game with specific learning goals [4], and commercial games where users can inhabit a virtual world with its own rules and learn or refine multiple skills by interacting with it [1]. Games that have been designed appropriately can be rigorously used for learning and assessment purposes [5]. First, games engage players in authentic environments that can portray versatile mechanics where users adapt to specific rules and constraints, facilitating authentic situations that resemble more realistically what they will encounter in real-life problems [6]. Moreover, the telemetry of games facilitates collecting rich data sets that, by applying analytics, can be used to reconstruct the entire problem-solving process instead of solely looking at the final outcome of the problem [7]. Finally, annual surveys on teen media use consistently indicate that playing games is a very popular leisure activity [8]; 90% of teens in the USA say they play some kind of video game (taking into account all devices). Therefore, people's positive experience of playing games in their daily lives can lead to better engagement when purposely designed games are introduced in educational contexts [9].

While games are constantly assessing players' performance [5], game-based assessment (GBA) uses educational games, their mechanics and tasks to generate evidence about skills and content knowledge that can be valuable [10], [11]. GBAs present unique opportunities for alternative assessment [12]. For example, because students see games as a leisure activity, they might feel less stressed compared to traditional testing methods [13]. Moreover, the assessment can be implemented unobtrusively without interrupting the game flow by using in-game measures as evidence that are fed to embedded assessment models [14]. Finally, because games elicit rich data about different processes and strategies for problem-solving, students can receive personalized feedback tailored to the specific challenges that they are facing [15].

Developing a well-balanced GBA that is both enjoyable and maintains high qualities of a good assessment is quite difficult [10]. It is particularly challenging to ensure psychometric qualities (e.g. validity), while managing high development costs, and addressing the technical challenges associated with developing the assessment machinery [16]. Because learning in games is interactive, dynamic, and multidimensional [17], many of the assumptions that traditional psychometric models require (for example, that latent variables stay constant for item response theory) can be easily violated. Therefore, developing robust yet flexible assessment models is one of the most challenging steps of GBA development.

In this paper, we focus on this key point of developing the algorithmic assessment machinery of a GBA. In the literature, we find multiple approaches to conduct the assessment such as rule models [18], performance indicators [19], or learner modeling using knowledge inference algorithms [10], [20]–[24]. Our work is focused on this last approach that uses

Manuscript received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY. This material is based upon work supported by the National Science Foundation under Grant EAGER #1935450. This work has also been partially supported by the MIT-SPAIN "la Caixa" Foundation SEED FUND. *(Corresponding author: José A. Ruipérez-Valiente.)*

José A. Ruipérez-Valiente and Pedro A. Martínez are with the Department of Information and Communication Engineering, University of Murcia, Campus Universitario, Murcia 30100, Spain (e-mail: jruiperez@um.es and pedroantonio.martinezs@um.es).

Yoon Jeon Kim is with the Department of Curriculum and Instruction, UW–Madison, 1000 Bascom Mall, Madison, WI 53706, USA (e-mail: yj.kim@wisc.edu).

Ryan S. Baker is with the Graduate School of Education, University of Pennsylvania, 3700 Walnut St, Philadelphia, PA 19104, USA (e-mail: ryanshaunbaker@gmail.com).

Grace C Lin is with the MIT Education Arcade, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: gcl@mit.edu).

Digital Object Identifier InsertFinalDOIHere

knowledge inference algorithms. We find multiple algorithms in the literature, each with their own potential benefits and drawbacks: Bayesian networks [10], [21], fuzzy cognitive maps [22], item response theory [23], or multi-dimensional factor modeling with DINA [24], among many others. In this paper, we discuss the potential of an adapted version of the Elo rating system [25] to perform multivariate learner modeling [26] for GBA. While Elo has been used over the last years for adaptive learning practices across several contexts [27]–[29], as far as we know, it is the first time that this approach has been leveraged for GBA purposes.

The overarching research question of this study is about discovering the benefits that multivariate Elo can have as part of the core game-based assessment machinery. We argue that multivariate Elo can present numerous advantages, especially considering the foundations of game environments in education [30], as it can lead to an accurate learner modeling performance without the need of accumulating large amounts of data [29]. Moreover, this algorithm is also useful to estimate task difficulty, which is always important to keep at the right level in games in order to maintain a good playing flow [31], [32]. We will explore these advantages through a case study with a geometry GBA with puzzle solving mechanics called Shadowspect. More specifically, we will demonstrate the potential of multivariate Elo-based learner modeling in games for the following objectives:

- To measure students competence simultaneously across multiple knowledge components in a game.
- To predict whether a student will correctly solve a task in the game.
- To estimate task difficulty within the game.

The remainder of the paper has the following sections. Section II reviews the background focusing on GBA and knowledge inference algorithms. Section III describes the materials and the methods applied while Section IV presents the results achieved. We finalize with a discussion in Section V and conclusions in Section VI.

II. Background

A. Game-based Assessment

In the past decade, the field of game-based learning has accumulated abundant evidence about how games can be successfully used to support learning academic content [33], to demonstrate the effects of prosocial behaviors [34], and to change attitudes and beliefs thus increasing civic participation [35], to name some examples. GBA is a specific application of learning games, referring to a type of assessment that uses players' interactions with the game, both digital and non-digital, as a source of evidence to make meaningful inferences about what players know and can do (i.e. knowledge, skills), and how individual players interact with the game as a problem-solving process [16]. Some games are developed specifically for assessment purposes by considering predetermined learning goals and elicited evidence for the game mechanic during the design process. However, the assessment machinery–a computational model that connects evidence with latent variables (i.e. psychometric model or developing learning analytics and automated detectors)– can also be retrospectively developed after the game is fully implemented [36]. GBA, like any other type of assessment, needs to meet certain psychometric qualities such as validity, reliability, and fairness [37].

In general, there are three approaches to assess learning in game-based learning. The first approach relies on administering external measures and instruments such as survey and observations [38]. One good example of such an approach is having players take questionnaires before and after playing a game. While commonly used, this is not truly GBA because inferences about players or learners are not directly relying on the evidence generated from game play itself. The second approach, becoming more common with the advancement of data science, uses in-game behaviors and indicators for performance to predict learning after playing the game [39]. Most GBAs fall into this category, where the goal is to evaluate if and how the players are learning from playing the game, rather than making assessment inferences about individual players. For example, Ruipérez-Valiente and colleagues [40] assessed different patterns of engagement in an online multiplayer game that targets middle school science inquiry practices. The third approach is a form of performance-based assessment, where a game is developed with specific assessment goals. These games include explicit alignment of game mechanics and features with assessment mechanics that are linked to the educational goals the designer wishes to make inferences about. Unlike the other type of GBA, this approach requires a computational model (e.g. algorithms, psychometric models) that enables evidence-based reasoning [41], which we call "assessment machinery" based on the evidence-based assessment framework [42]. These models can be embedded in the game engine to make assessment activities seamlessly woven into the game play itself. This approach, also known as stealth assessment [14], is the strictest form of GBA, as the student never notices the assessment side of the game. For example, Physics Playground [37], a simple-machine based game developed with the goal of assessing middle school students' conceptual understanding of physics, is an example of this type of GBA. There have been multiple approaches to implementing the assessment machinery in games: using experts to do human labeling and then conducting machine learning, knowledge engineering, rule models, and knowledge inference algorithms. We focus on the last approach and elaborate on this in the next subsection.

B. Knowledge Inference

Knowledge inference is a key component of many contemporary educational technologies [43]. Knowledge inference algorithms can be used for multiple objectives, but the most common use is modeling learners' knowledge in order to perform adaptive learning, which may include assessing whether a student has become proficient enough to move to the next subject [44], to select the next topics or items of a student [45], to adjust difficulty [46], and many other uses. Knowledge inference models are also frequently used to provide reports of student skill to teachers [47]. One important framework

defined within this context has been the Knowledge-Learning-Instruction framework (KLI) [48], which relates observable (instructional and assessment events) and unobservable events (learning events); these learning events are connected to the knowledge components (KCs) of students, which is how we structure knowledge modeling in our work.

One important framework defined within this context has been the Knowledge-Learning-Instruction framework (KLI) [47], which relates observable (instructional and assessment events) and unobservable events (learning events); these learning events are connected to the knowledge components (KCs) of students, which is how we structure knowledge modeling in our work.

Inferring student knowledge in an interactive environment such as games is a different challenge from inferring student knowledge in a test or standardized examination for the simple reason that students are learning at the same time as their knowledge is being measured. In other words, knowledge is changing while it is being assessed.

The most widely-used algorithm for this purpose (at least in terms of application within educational technologies of various sorts) is bayesian knowledge tracing (BKT) [44]. Relatively straightforward to apply, and relatively straightforward in its behavior and applications, BKT assumes a simple two-state Markov Model where students transition from not knowing a skill to knowing a skill, with data-fit probabilities for guessing and slipping. BKT has proven useful for a range of research applications as well as educational use [49]–[51]. However, BKT tends to perform more poorly than more recent algorithms in terms of ability to forecast future performance within the learning tool [52], [53] and external to it [54].

The second most widely-used algorithm within running learning systems is Elo [55]. First developed in the context of Chess rankings, Elo became widely used in adaptive learning systems within Europe starting around a decade ago [28]. Similar mathematically to item response theory (but possible to use in real-time with changing knowledge), Elo postulates that each student has an ability level, each item has a difficulty, and estimates continually adjust based on ongoing student performance.

A third popular family of algorithms is the logistic function based performance factors analysis (PFA) [56] and its several modern extensions [57]–[59]. PFA predicts performance (and more indirectly infers knowledge) based on a student's past proportion of correct and incorrect answers, as well as overall skill or item difficulty. PFA can be somewhat slow to recognize that a student's ability has changed if they have considerable amounts of unsuccessful work; more recent extensions weight recent actions higher in estimation [57]–[59].

Finally, recent work has seen a proliferation of algorithms for knowledge inference based around neural networks [52], which can find highly-complex relationships between items. These algorithms tend to do considerably better than older algorithms at predicting future student performance within learning systems [53]. However, they retain some limitations for use in real-world educational systems: many variants do not address concerns raised early in the development of this type of algorithm, that estimates can be unstable and counter-intuitive (predictions going down after a correct answer) [60]; these algorithms do not generally provide estimates of teacher-interpretable skills, although recent extensions provide a possible path forward [54]; and much of the difference in prediction quality may involve prediction of performance before a student even begins a skill rather than differentiating between students with considerable practice [61]. As such, more work is still needed before these algorithms can be widely used to provide reports for teachers.

Most of the work on knowledge inference has been applied in the context of intelligent tutoring systems which have well-defined and less open-ended tasks [62]. As learning environments become more open in design such as games – considering a broader range of behavior to be correct for the purposes of knowledge inference– the design of knowledge estimation becomes more complex. For instance, a range of behaviors can be recognized as correct within the Inq-ITS science simulation learning environment, requiring a two-step process where first an algorithm is applied to identify the probability that the behavior on the current simulation is correct, and then its output is fed into a knowledge inference algorithm as a second step [63]. This example highlights that as more open-ended activities such as learning games and simulations need to recognize a wider range of behaviors as correct (or partially correct) [64], knowledge inference needs to be adjusted correspondingly.

After this brief introduction, we will now describe Shadowspect, its context, and why and how we implemented multivariate Elo-based learner modeling for the assessment machinery of geometry standards.

III. Materials and Methods

A. Shadowspect Game

Shadowspect (see a video online[1] was designed as a GBA tool with the objective of generating metrics about geometry content as well as other constructs related to their behavior and cognitive skills. Shadowspect was designed the goal of providing teachers' with a formative tool to teach and evaluate Common Core Geometry Standards (e.g. visualize relationships between 2D and 3D objects) [65], enabling teachers to use the tool within their math curriculum activities. In this paper, we describe the process that we followed to develop the assessment machinery based on Elo learner modeling algorithm.

Figure 1 depicts a couple of the existing puzzles in Shadowspect. We can see red parallelograms in both puzzles, that are delimiting meaningful areas that we will explain next. Once a student starts a puzzle, they can read a brief description of the task (A) and they receive a set of silhouettes obtained from different viewpoints that depict the final figure they have to construct (B). To solve this challenge, students can generate the following primitive shapes: spheres, cylinders, pyramids, cubes, ramps, and cones (C). Moreover, certain puzzles could also have additional constraints, for example, a maximum number of shapes or having only certain shape types available.

[1] https://youtu.be/j1w_bOvFNzM

Then, learners can use available tools for moving, rotating, and scaling shapes, in order to build a figure that would match all the silhouette viewpoints provided by the puzzle (D). In addition, they can select or delete multiple shapes with a single action. There are other important actions, students can also modify the camera view to observe the figure they are constructing from a different viewpoint (E) as well as use the 'snapshot' button to produce the silhouette from the actual viewpoint (F). These snapshots are useful to help the student know if the current figure is matching some of the silhouettes of the solution. Finally, the student can submit the current figure and the system will assess if the current figure is correct, providing the student with automatic feedback (G).

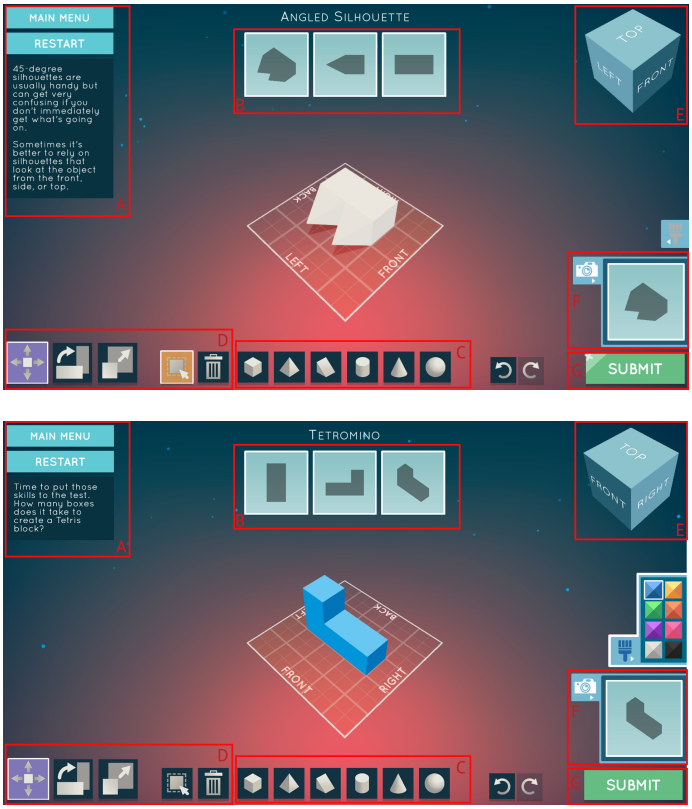


Fig. 1: A couple of the existing puzzles in Shadowspect

B. Puzzles and Knowledge Components

We initially designed a total of 30 puzzle levels in Shadowspect, where we categorize nine of them as tutorial, nine as intermediate, and 12 as advanced. The tutorial levels just focus on facilitating the understanding of the game mechanics, so that students learn all the functionalities, like creating or rotating shapes, changing the perspective, or making snapshots. Students receive direct help in the game to overcome the tutorial. Students receive little to no scaffolding in the intermediate levels, while the advanced levels are meant to be challenging for students who have gained experience with the previous levels.

This set of puzzles in Shadowspect were collaboratively designed by a game designer and a learning scientist; a math teacher consulted on the design process to identify and align the Common Core Standards to the puzzle designs. We focused on the Common Core Standards at the level of 6–9th grades [65]. The teacher reviewed each of the geometry Shadowspect puzzles and coded which of the Common Core Standards were present. We defined three potential codes, which were 'none' in the case that were was no relationship between the puzzle and the standard, 'weak' in the case that the puzzle provided weak evidence of the standard, and 'strong' in the case that the puzzle provided strong evidence of the standard. After developing the coding process, we identified four standards and we provide next their exact descriptions from the official site of the Common Core State Standards [65]:

- `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."

Within the context of knowledge inference, knowledge is often considered in terms of KCs, which are associated with every problem-solving item, which in the case of Shadowspect are the puzzles. KCs were defined within the KLI framework as acquired units of cognitive function or structures that can be inferred from performance on a set of related tasks [48]. Therefore, within our context we will define each one of Common Core Standards as a KC. In Table I we depict the results of the full mapping between puzzles and KCs.

C. Context and Data Collection

Shadowspect was designed as a GBA that could be used across several high school grades, and it has been tested in multiple studies and settings. For this paper, we utilize data from an original data collection that was performed to build the assessment machinery of Shadowspect. The development team recruited seven teachers across the USA to use Shadowspect for at least two hours in their seventh to tenth grade math and geometry classes. The use of the tool was instructor-paced, meaning that each instructor selected which days of class Shadowspect would be used. Across these teachers' classes, there were no limitations in terms of the number of trials that students could complete on a puzzle and students were not required to play for a minimum time during the sessions. We provided support to the teachers to perform the data collection by generating personalized URLs for their classes, solving technical doubts, and pedagogically helping with the use of the tool. Data was collected from 322 different students.

TABLE I: This table denotes the full mapping of each puzzle to the detected KCs. The symbol ✗ indicates no evidence, the symbol ↗ signals a weak evidence, and the symbol ✔ a strong evidence.

Puzzle	MG.1	GMD.4	CO.5	CO.6
Squared Cross-Sections	✗	↗	↗	↗
Bird Fez	✔	↗	↗	↗
Pi Henge	✔	↗	↗	↗
45-Degree Rotations	✗	↗	✔	↗
Strange Pyramids	✗	↗	↗	↗
Boxes Obscure Spheres	✗	↗	✔	↗
Object Limits	✗	↗	↗	↗
Not Bird	✗	↗	↗	✔
Angled Silhouette	✗	↗	↗	↗
Warm Up	✗	↗	↗	↗
Tetromino	✗	↗	↗	↗
Stranger Shapes	✗	✔	↗	↗
Sugar Cones	✗	↗	↗	↗
Tall and Small	✗	↗	↗	✔
Ramp Up & Can It	✗	↗	↗	✔
More Than Meets Your Eye	✗	↗	✔	↗
Unnecessary	✗	↗	↗	↗
Zzz	✗	↗	↗	↗
Bull Market	✔	↗	↗	↗
Few Clues	✗	✔	↗	↗
Orange Dance	✗	↗	↗	↗
Bear Market	✗	↗	↗	↗

The game was developed as a Unity application, hosted as a web application, and all interactions of students with the game are recorded in the database, enabling the reconstruction of the learning process that students performed to resolve any puzzle. By design, we decided to not collect any private information from students, except for a nickname that they selected themselves.

We collected data from 322 students building a data collection with approximately 428,000 gameplay records, which represents around 1,320 events per user. Students played 260 hours and solved 3,802 puzzles, which implies that the average student was active for 0.82 hours and solved 13 puzzles. Note that even though teachers conducted two hour sessions with Shadowspect, the total active time (i.e., actively engaging with the game) of students within the game environment was in most cases below two hours as students exhibited different degrees of engagement with the game.

D. Adapting the Elo Algorithm to Multivariate Elo-based Learner Modeling in the Shadowspect Game Environment

Multivariate Elo has been chosen as the knowledge inference algorithm in Shadowspect for several reasons. First, we wanted an algorithm that accounts for puzzle difficulties while predicting performance of the overall proficiency on four KCs within the Common Core Geometry Standards. Second, Elo is a relatively simple algorithm for both implementation and interpretation. Given that Shadowpect is designed as a formative assessment tool for teachers, the high interpretability was a priority for which algorithm should be chosen. Third, given a data set, the algorithm is fast to converge into good-performing estimates of the item difficulty and students' competency and can do so with a limited data sample in terms of size [29]. Moreover, if the algorithm is introduced in a different context where there are new students or items, it will also generate those estimates properly in few steps. Lastly, Elo has reasonably good predictive performance in terms of classification metrics, like AUC, when compared to other state of the art knowledge inference algorithms such as Bayesian networks [29]. However, we discovered that Elo has some challenges for application in this context, which we will return to later.

The rest of the section is divided in two parts. The first depicts the basic foundations of the Elo algorithm proposed by Arpad Elo [25]. The second explains how we adapted the algorithm to perform multivariate Elo-based learner modeling within the specific context of Shadowspect.

1) Basic Elo Algorithm: The Elo algorithm was first (and is still) used as a method to rank players in chess and across other games in order to analyze the performance in a match between two opponents. The basis of the algorithm is that the performance of each player is a random variable that follows a bell-shaped curve normally distributed over time. After each game, the values of both players are updated [25]. If the player with the higher ranking value wins, a minor update is performed. Otherwise, the performance values of both players undergo a larger change.

By analyzing the estimated ability values of each player, we can predict the likelihood that each player has of winning the game. The formulas to calculate the expected score of player A and player B are [25]:

$$EA = \frac{1}{1 + 10^{(RB-RA)/400}} \tag{1}$$

$$EB = \frac{1}{1 + 10^{(RA-RB)/400}} \tag{2}$$

Where RA is the performance of player A and RB is the performance of player B. These are the foundations of Elo algorithm that are applied for ranking players based on their Elo scores. However, these foundations can also be applied within educational contexts to estimate the competency of a learner and the difficulty of the items in a learning environment [55]. In our context, each of the puzzles in Shadowspect is considered an item. The main conceptual difference with respect to the basic Elo ranking algorithm is that in learner modeling we confront a student with an item, which takes the form of a puzzle in Shadowspect, and instead of inputting two players' statistics into the logistic function, we input the estimated competency of the student and the estimated difficulty of the puzzle. Following the same foundations as before, when a student with a low competency solves a hard puzzle, the estimate of the competency of the student will increase significantly and the estimated difficulty of the puzzle will decrease. In the opposite case, if a high competency student fails to solve an easy puzzle, the estimate of the proficiency of the student will decrease and the estimated difficulty of the puzzle will increase. By applying this Elo-based learner modeling, we can estimate the competencies of learners and the difficulty of the items on a single KC. Below we explain the formulae to extend it to multiple KCs.

2) Multivariate Elo-based Learner Modeling: First, we need to compute how probable is for a particular student to solve an item correctly. Then, given a student s that is confronting a item i, the algorithm uses a logistic curve function that incorporates both the competence of the student θ_s and the difficulty of the item d_i:

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

To compute this probability, θ_s and d_i are respectively the average competency and difficulty of all the KCs that are present in item i (in our case scenario this mapping is depicted in Table I). The following formulas will be necessary to update both the difficulty of the item and the competence of the student. In these formulas the α normalization factor is used to ensure the zero-sum principles in the model, which is calculated as follows:

$$\alpha_{den} = \alpha_{den} + (ans - P(\theta_s, d_i)) \tag{4}$$

$$\alpha = \frac{(P(\theta_s, d_i) - ans)}{\alpha_{den}} \tag{5}$$

Then, each step of the algorithm takes place when a student attempts to solve an item. At that point, to update the difficulty of the item, we need the calculated probability that student s has to solve item i ($P(\theta_s, d_i)$), if the response to the item ans was correct or not (variable ans takes a 1 if the response was correct and a 0 it was incorrect), the α normalization factor, the number of attempts to items n that student s has completed, and the γ and β adjustment hyperparameters. We assigned the values of γ and β to 1.8 and 0.05 respectively based on previous work that found through simulations that these values were stable in most settings [26]. Moreover, we also have a weighting variable w_k which is the weight that KC k has in puzzle i. Therefore, to compute the change in the difficulty of item i, for each one of the KCs (k) that are present in the item i we have that:

$$d_{i,k} = d_{i,k} + w_k * \frac{\gamma}{1 + \beta + n} * \alpha * (P(\theta_s, d_i) - ans) \tag{6}$$

In addition, to compute the change in the competency of the student s, for each one of the KCs (k) that are present in the item i we have that:

$$\theta_{s,k} = \theta_{s,k} + w_k * \frac{\gamma}{1 + \beta + n} * \alpha * (ans - P(\theta_s, d_i)) \tag{7}$$

With this approach, we can compute the change of the competency and the difficulty considering the importance of each KC in those items that contain more than one KC. That means that, if an item has two KCs, one providing a weak evidence and a second one providing a strong evidence, after each completed attempt, the change on the competency of the student and the difficulty of the item associated with the weak KC, will be smaller than on the competency and difficulty associated with the strong KC. With this approximation, we are able to have a flexible algorithms that accommodates items that have more than one KC present at the same time, and adjust the updates based on the strength of the evidence provided by each item. Moreover, the γ and β adjustment hyperparameters determine the initial value and the slope of change respectively, so they have an important influence on how big the changes are in the difficulty and competency. In many occasions, we need to find a balance between the speed of the convergence and the stability of the estimated parameters; meaning than a higher speed of convergence would also cause to have estimated parameters with a higher variance over time. This speed of the convergence can be an important factor in GBA, if the assessment machinery needs to be train across different contexts (different classroom ages for example) and with limited data sample sizes.

3) Design Decisions to Adapt it to Shadowspect: While the previous sections describe the general formulae of the multivariate Elo-based learner modeling, we had to make a number of decisions to accommodate the features of games as a learning environment. We anticipate that other GBA development teams will have to go through a similar process to use the multivariate Elo algorithm due to several reasons. Since games are much more open-ended environments than other learning systems such as intelligent tutors, it is harder to map students' behavior into the algorithm parameters, forcing the designer to make decisions in order to map the evidence into the algorithm. Moreover, GBAs should still feel like games; therefore, there is a tension between the assessment side and the fun side, making tasks more diverse and multifaceted depending on the specific context and area of application. We now report some of the decisions that we had to make within the context of Shadowspect GBA.

First, we decided that tutorial levels should not impact the estimation of student competency. This decision was based on the fact that these levels are guided puzzles where the primary objective is to facilitate users to learn the game mechanics. However, we do still compute the difficulty of such levels for game design reasons and to be able to retrieve that information if necessary.

Moreover, we also defined within the context of Shadowspect what we meant by an attempt, which is an important definition for knowledge modeling algorithms. One key aspect is the functionality of Shadowspect allows students to start and exit any of the existing levels without any restrictions. Additionally, each time a level is started, the student can make multiple submissions without exiting the level. Therefore, within this context we defined an attempt as all the actions performed between starting a level and exiting the level, whether the level was correctly solved or not. If the student enters a level, but never submits a solution, this attempt is discarded. If the student returns in the future to the same puzzle after a failed attempt, it will be considered as a new attempt on the puzzle as the student failed the previous one and, after attempting other puzzles, is trying this one again. However, if the puzzle was solved correctly in a previous attempt, a new attempt will not be taken into account as the student already knows the solution to the puzzle.

Finally, we provide a partial assignment approach based on the variable w_k that affects the change in the competency and

difficulty based on the evidence provided by the puzzle on that KC (none, weak or strong as depicted in Table I). In our case, there was either one or none KCs categorized as strong per puzzle, and each puzzle provided one unit of evidence. That is, if a puzzle provides strong evidence for a KC, w_k takes half the unit of evidence (i.e., 0.5) for that KC, and the other 0.5 unit of evidence is divided equally between the rest of weak KCs that are present. If the puzzle has no strong KC and multiple weak KCs, then the one unit of evidence is divided equally between the weak KCs present. There remains considerable debate about how to assign credit in cases like this one (see review in [66]); this approach was chosen in order to assign more credit to strong KCs but less to weaker KCs, without having to distinguish relative levels of contribution beyond strong/weak. It also would have been possible to fit parameters for relative contributions (as in [67]), but doing so would not have changed the overall goal of the approach, and achieving the best possible fit is not the goal of this paper. In addition, not fitting parameters for relative contributions increases the potential for this approach to rapidly generalize to new content. These decisions demonstrate the subtleties of adjusting the Elo algorithm to a specific game and context. Other projects may wish to fit parameters or select a different approach to credit assignment based on the specific tool, context, and goals of the modeling approach.

IV. Results

In this section we present the results of each one of our three original objectives, namely the measurements of students competences (Subsection IV-A), the task performance prediction (Subsection IV-B), and the item difficulty estimation (Subsection IV-C).

A. Student Competence per Knowledge Component

After applying the previously described methodology and multivariate Elo-based learner modeling algorithm, we have the estimated competencies of each student in each one of the KCs to demonstrate our first objective. Figure 2 shows the competency distribution per KC for a selection of the students in our population. The competencies are displayed with a stacked bar chart where each color represents a KC. The competency for each KC is normalized from 0 to 1, where 0 would be the lowest value of the KC in our student population and 1 would be the highest. We see quite a diverse distribution of competencies within this sample of students, even though the teachers across different classes encouraged students to play only for two one-hour sessions. Some students might have a higher value in one competency than in others depending on their skill level. However, this difference is not very large, in part due to the substantial overlapping and the distribution of the KCs per puzzle, as we saw in Table I.

The capacity of the current setup and model to differentiate between the four KCs can be quantified based on the correlation between them, which are presented in Table II. As we can see, there is a very high correlation, above 0.95, between GMD.4, CO.5, and CO.6 KCs that suggests that these clearly form a single factor, whereas MG.1 has a low correlation of around 0.14 with the rest of the KCs. This overly-high correlation between most KCs has been caused by the top-down approach that we followed by first designing the puzzles, then coding existing KCs of each puzzle. The other potential alternative would be performing a bottom-up approach by first specifying the desired KCs at the beginning of the design process, then designing the puzzles that best fit into that criteria.

TABLE II: Correlation matrix between the four KCs.

KC	GMD.4	CO.5	CO.6	MG.1
GMD.4	1.00	0.97	0.98	0.13
CO.5	0.97	1.00	0.98	0.14
CO.6	0.98	0.98	1.00	0.15
MG.1	0.13	0.14	0.15	1.00

Furthermore, we can also analyze the evolution of the competency of each student after each of the attempts to a puzzle. In the Figure 3 we see the value of the competency of one selected student after each one of the puzzle attempts, which are represented in the *x*-axis. The green color denotes a successful attempt, while the red color implies a failed attempt. We can see that the first initial puzzles were correct, but there was no change in the competency. This is due to the fact that those are the nine tutorial puzzles which, as we explained in our implementation description in the methodology section, we do not take into account for competency estimation. However, the difficulty of these puzzles is still inferred (as we will see in the next subsection). We can see in this case the student solves a number of puzzles correctly in a row, from 'Square Cross-Sections' to 'Sugar Cones,' but then failed in their attempt on several subsequent puzzles which led to a decrease in the competency estimate, specifically in GMD.4.

The competency estimation can be used as part of the formative assessment process with the GBA tools, and the evolution over time is useful for teachers to perform just-in-time interventions based on the current status of the student.

B. Task Performance Prediction

The second objective was to demonstrate how this Elo algorithm can be used to accurately predict the outcome of each one of the puzzle attempts based on the estimated competencies and difficulties. To do so, we report in Table III some common classifier metrics for 2-class binary problems, where all of them are computed for the correct class i.e. '1'. These metrics are computed based on the prediction that the algorithm makes for each one of the puzzle attempts that a student performs. We can see that the algorithm makes a correct prediction 94% of the time, has area under the ROC curve of 0.87 (the model's ability to distinguish correct attempts from incorrect attempts), and an F1-score of 0.97 (the model's ability to balance precision with the ability to recognize all cases), which can be considered very high values for a 2-class binary classification model (see also [53]).

The high performing metrics obtained demonstrate that we can use the model to accurately predict the performance of a student in a puzzle. This model could consequently be used for adaptive learning purposes, for example by adapting

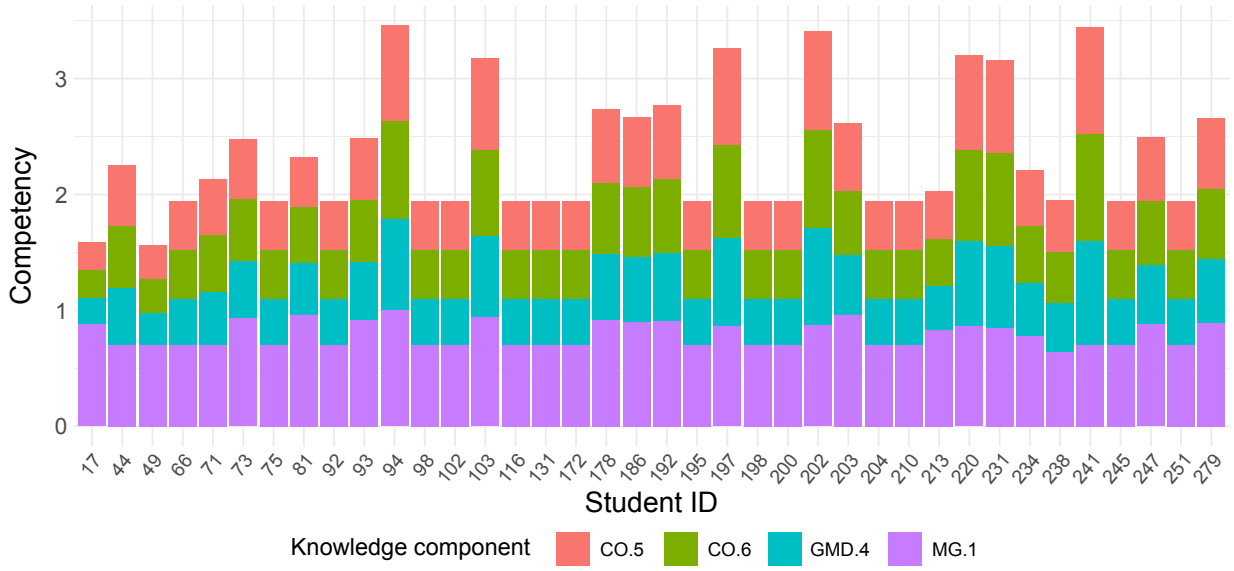


Fig. 2: Stacked bar chart with the competency of each student divided per KC for a selected sample of students in our population.

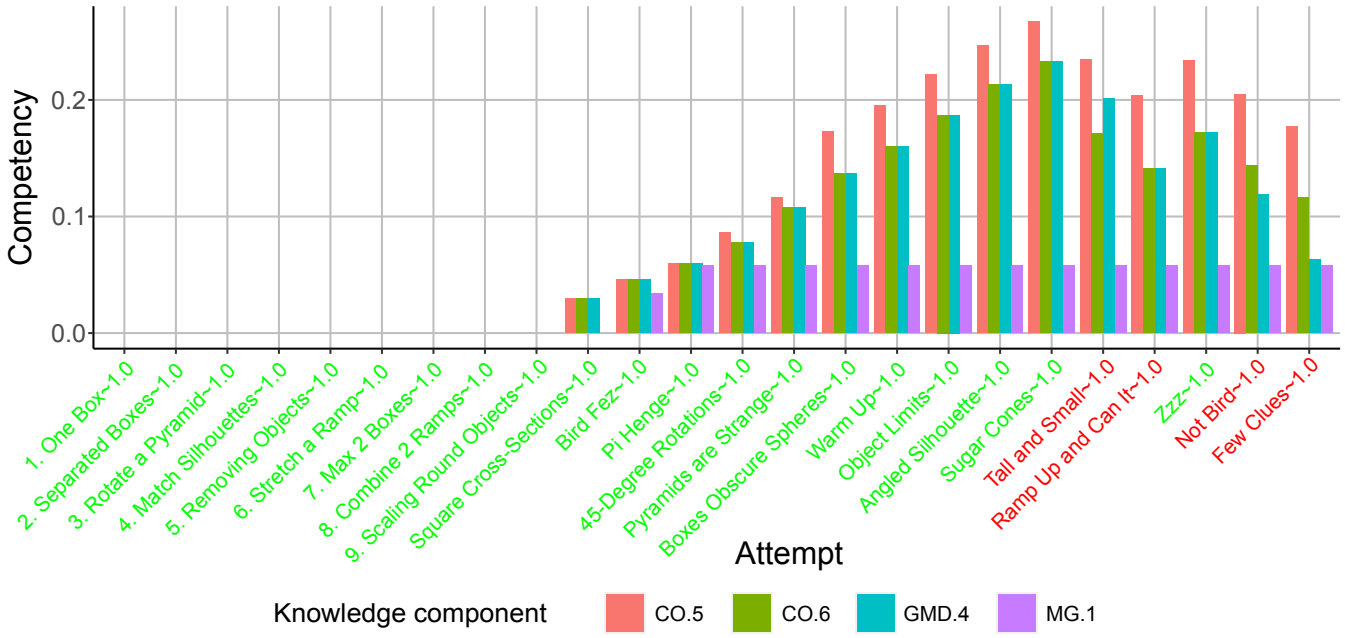


Fig. 3: Evolution of the competency of a selected student per KC attempt per attempt. Each attempt is represent on the x-axis, where the green color denotes a successful attempt to that puzzle and the red color a failure.

the sequence of puzzles of a student based on their current competency level in order to keep them in the zone of proximal development [68]. After a student completes each puzzle, the algorithm could find a puzzle of appropriate difficulty across KCs based on the student's current competency estimates.

TABLE III: Classifier metrics based on the predictions of each attempt by the multivariate Elo-based learner modeling algorithm.

Accuracy	AUC	F1-score
0.94	0.87	0.97

C. Item Difficulty

The last objective was to show how multivariate Elo-based learner modeling can also be used to estimate the difficulty of the items in a GBA environment. In our case scenario, we will show the difficulty estimation of each one of our puzzles. Figure 4 shows the difficulty per KC in a stacked bar chart

for each one of the puzzles that we have in Shadowspect. The difficulty depicted is separated by KC, but it is also possible to compute an averaged difficulty by considering the importance of each KC in a puzzle – effectively what our Elo variant does when predicting student performance.

Moreover, we can also validate the difficulty estimated by Elo by correlating it with a more deterministic difficulty measure that we generated as part of previous work in [31]. In this previous work, we computed a knowledge engineered difficulty metric based on the time invested, the number of actions, the percentage of incorrect attempts, and the attrition rate of a puzzle averaged across all the students that attempted each puzzle. If we compute a correlation between this knowledge engineered difficulty metric and Elo difficulty metric we obtain a correlation of 0.76. This is quite a high correlation, especially when we take into account that the two measures of difficulty are computed following a intrinsically different procedure.

This difficulty can be used as part of the GBA design process and in order to better understand the success or failure of students with certain easy or hard puzzles.

V. Discussion

The discussion is divided in an initial part that discusses the implications of the results for GBA (Subsection V-A) and a second part that shares some insights regarding the algorithmic design (Subsection V-B).

A. Implications for GBA

We believe these these results may have implications for the implementation of assessment machinery of future GBAs. More specifically, for this study we established three objectives to demonstrate the affordances that multivariate Elo-based learner modeling offers within the context of GBA.

First, we have adapted the Elo ranking algorithm for learner modeling while accommodating multiple KCs using a multivariate approach. The Elo algorithm was historically used for player ranking by facing one player to another, but in the learner modeling variant we face one student with an item, in our case scenario each of the Shadowspect puzzles are the items. While using Elo for learner modeling is not novel in the literature [27], [28], [55], to the best of our knowledge this is the first time that it has been applied within the context of GBA. The proposed algorithm presents several strengths that make it a great option for its use in the context of games as assessments. For example, the proposed algorithm allows for items (or puzzles in this case) with multiple KCs, and it is quite easy to fit competency and difficulty parameters that are both meaningful and interpretable. This flexibility is crucial for a GBA when the underlying competency model is likely multidimensional [69]. It also allows adding weights (partial assignments) for each one of the KCs in an item, reflecting the designer's understanding of how each item aligns with the KCs. In our case, we consulted with a geometry teacher, and mapped each one of the puzzles in Shadowspect to the associated KCs based on the evidence that the puzzle provides [10].

Second, although we did not implement this in the current work, this algorithm can be used in GBA to predict the potential performance of a student when facing a new game task. We evaluated the quality of the learner modeling via classifier metrics and obtained good model performance results with an accuracy of 0.94, AUC of 0.87, and F1 of 0.97. This shows that it will be viable to use this model to predict the likelihood that a student will solve a level correctly or use that prediction to select which level a student should work on next. [26]. This would help both to keep users in the state of flow and in the zone of proximal development [31], [32], thus improving the learning-difficulty balance by having an adapted experience of the learner with the game. These adaptive game-based assessments could more autonomously and rapidly adjust players into puzzles with more suitable difficulty levels by stealthily assessing their knowledge and adapting their sequence of tasks. This kind of system would have similar characteristics to adaptive intelligent tutoring systems [70]. Furthermore, the algorithm presents advantages for use in these contexts as it can converge quickly to reasonably good difficulty and student competence estimates, which help with cold-start issues when either new puzzles or students are introduced into the system [71].

However, the implementation of adaptivity features would need to be designed carefully, as these are systems designed for assessment purposes, and therefore we need to guarantee the validity of the assessments and also that all players are assessed in a equitably fair way [37]; therefore, adapting the tasks of players differently might impact the reliability of such assessment properties. These are challenges that have been addressed in past work on adaptive assessment, and they are definitely possible to address here, but careful effort will be needed to do so. Moreover, previous work in the context of game-based learning adaptivity has in many cases reported null results, indicating no differences between adaptive level navigation and linear student navigation [72], [73]. In our specific case, the value of adaptation in Shadowspect puzzle selection has not yet been experimentally demonstrated, and thus it remains as a direction for future work.

Third, this algorithm can work well for GBA because of its capacity to quickly generate data-driven estimates of the game level difficulties. Inferring the difficulty of the game levels is an important aspect in game design in general [74], but it is particularly important within the context of educational games [31] to avoid learners feeling that the game is either too hard or easy, as this would impact negatively the learning experience, so the objective is to maintain them in the flow zone [68]. Because the algorithm converges to good difficulty estimates quickly with a limited number of attempts, it can be easily used as part of the common play testing sessions performed within game development process [75], in order to have an idea of the difficulty of the tasks during based on the collected data. Moreover, we have validated this Elo-driven measure of difficulty by comparing it with a knowledge engineered difficulty that we computed in previous work [31]. We obtained a correlation of 0.76 between the two difficulty measures, which means that the difficulty estimates generated by Elo algorithm are similar to a carefully knowledge-engineered measure developed based

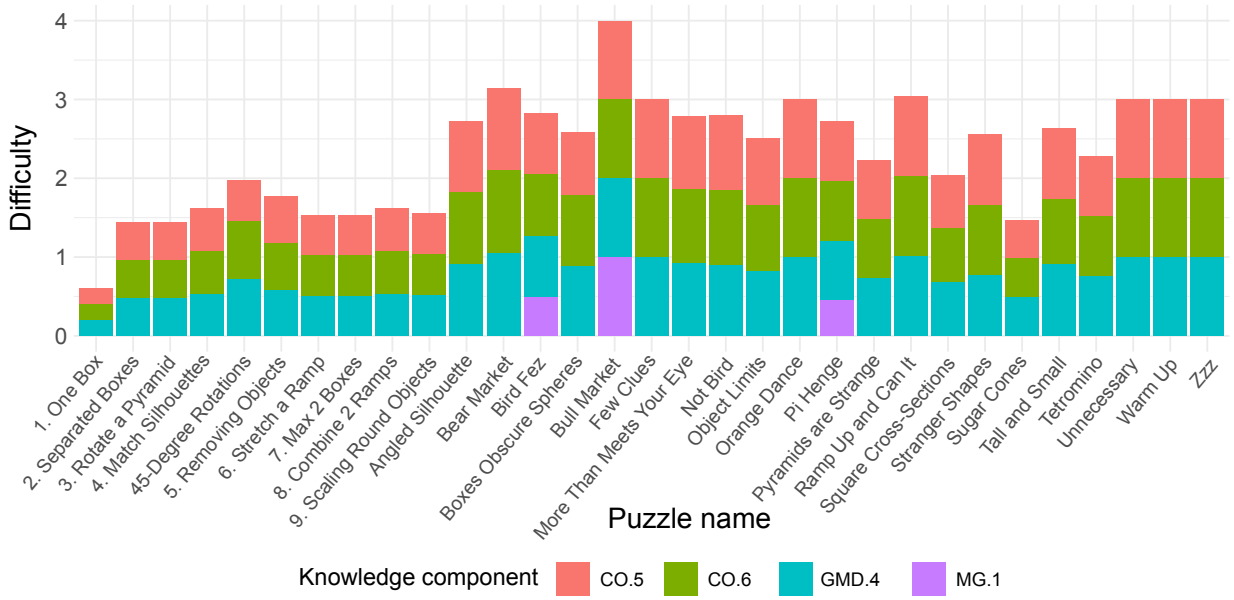


Fig. 4: Estimated difficulty for each one of the puzzles in the x-axis separated by KC.

on an extensive knowledge of difficulty in this game.

One of the main limitations of our work is that we did not obtain evidence for external validity [20] yet, for instance, through an external measures of student knowledge. The objective of such validation would be to confirm that the measures estimated through the multivariate Elo algorithm are indeed correlated with external measures of the geometry Common Core Standards. Moreover, we have applied this algorithmic methodology to one game environment only, therefore our lessons learned might not generalize across all game environments. We believe that this approach can be applied to other game environments which are clearly organized into independent levels, where the evidence generated in those levels can be mapped to multiple KCs. Fully establishing this will depend on obtaining additional data from other games. One final intrinsic limitation of multivariate Elo in its current form (and indeed of most of the algorithms used for student knowledge modeling) is that even though we collect rich data from the game, we only use the "right" and "wrong" information from the submitted items.

In summary, we believe that multivariate Elo-based learning modeling represents a strong option for the algorithmic machinery of GBA. To the best of our knowledge, this is the first study applying Elo within GBA. Future work should test this algorithm in other GBA contexts.

B. Algorithmic Design Insights

The design and implementation process raised many lessons learned that would be helpful for researchers attempting to implement this algorithmic approach in their GBAs (or other interactive learning environments). We will share those in this section.

Knowledge inference algorithms are often implemented in intelligent tutoring systems and other kinds of smart learning environments where the tasks or exercises that students need to solve are much more constrained in terms of what students can do in them. However, games normally have environments that are more open and where students have a higher degree of freedom to perform multiple actions that might be unrelated to the KCs. There might be design decisions that would normally be quite simple in a learning environment where the student has a very limited number of actions available (e.g. the student might just be able to provide answer to questions when solving a problem), but in the game environment, many decisions would require to thoroughly take into account the mechanics of the game. For example, in [29], Elo-based learner modeling was incorporated into a intelligent system to help students learn geography, where the system just uses two simple type of questions regarding a selected place in a geography map. Students can either get each question correct or not, and after that submission they move to the following question. This setup is quite convenient for knowledge inference algorithms and can be adapted with ease. However, in the case of Shadowspect, students can carry out multiple actions in each level (such as changing the color of shapes, doing snapshots, removing all elements...), submit multiple times receiving feedback on the correctness of their solution, and even come back to a level that they already solved correctly. This openness requires a thorough understanding of the game and the existing evidence of the competencies the designer is trying to infer to make a proper mapping. Therefore, the development of the assessment machinery might need the collaboration of the game designer, analytics modeler, and subject matter expert (in our case a geometry consultant).

Another important aspect involves the design of the game levels and the mapping of each level to the KCs that we want to infer. As part of the development phase of Shadowspect, the game designer built a number of levels without clearly targeting specific KCs. Our final result after retrospectively mapping the KCs to each puzzle is that there was a significant

overlap between them. This is demonstrated by the high correlation of 0.95 that we have in the results between GMD.4, CO.5, and CO.6. These high correlations were obtained even after taking into account domain expert perspectives on how strongly each puzzle was associated with each KC. This result indicates the value of performing a principled design of the game levels by, for example, conducting domain modeling using evidence-centered design (ECD) [42] where the design team identifies the needed KCs as a first step and then designs levels specifically to assess the target KCs in a focused fashion [76], one clear primary KC per level. In our case, each puzzle was mapped to a single Common Core Standard with the strong evidence category, but the same puzzle could also be mapped to other standards with the weak evidence category. The authors took this decision with the math consultant to increase the difference in the competencies between the strongest standard of each puzzle, and other weaker standards in the same puzzle. A different decision would impact the weights per standard when updating the competency of the student after a submitted response.

We also tried several options for the training of the algorithm that aligned with different pedagogical perspectives. For example, we made a two-run version, where the first run of the algorithm would be used to estimate the difficulty of the items, and the second run would be used to estimate the competency of the students while keeping the difficulty of the items fixed. This process supports teaching scenarios where a teacher wishes to directly receive the competency scores and intends to use them to compare students. If difficulty is not fixed in this instance, the fluctuation of the difficulty could negatively impact the effect that solving puzzles correctly or failing has on the competency model for different students. This could cause confusion to the teacher, and perhaps also an unfair situation for students if these scores are to be used as any kind of assessment. Alternatively, it is possible to pre-train the model estimating the difficulty of the levels in one sample of students, and then use those difficulty values for a new sample of students, which would help the model to converge much faster. Adjusting the value of the γ and β hyperparameters can make the convergence go slower or faster [27], [55], but this needs to be done very carefully to avoid undesired behaviors like an unstable convergence and weighting too heavily the last few attempts [55]. Lastly, since a student can retry the same puzzle level many times after failing in Shadowspect, we also considered the possibility to adjust the penalty after a failed attempt based on the number of previous attempts. This would help to decrease the penalty of failure to focus more on the final puzzle outcome. All these subtleties can help make the implementation of the multivariate Elo-based learner modeling more successful in GBA contexts.

VI. Conclusions and Future Work

Although teachers may be excited about the potential use of game-based learning and assessment approaches [77], the actual use of games in classrooms is often constrained by systematic requirements (e.g. [78]), the tension between playfulness and accountability, and teachers' own data and digital literacy. Therefore, we need to develop reliable and valid assessment models that can serve as a trustworthy source of useful information for teachers. This study has highlighted the potential that multivariate Elo-based learner modeling has within the context of GBA in order to estimate the competency of students in multiple KCs at the same time, to predict task performance in each attempt, and to estimate task difficulty in the game. We have done so with a case study using Shadowspect, a geometry GBA based on 3D puzzle mechanics. Therefore, we encourage future researchers in the field to consider this algorithm to build their assessment machinery.

There is still future work open in several directions, for example to validate the models' assessments to external measures (in our case, perhaps, a standardized test of the geometry Common Core Standards). In doing so, it would be valuable to compare the performance of this algorithm with the one of other more commonly used GBA algorithms like BKT, IRT or PFA. Moreover, since the final objective is to introduce these tools into the classroom, future work should work on increasing the interpretability and trustworthiness of these competency scores from the teachers' side. In doing this, it would be essential to develop dashboard designs [79] that allow teachers to explore the concepts, dig into the related indicators that feed into the scores, and construct their own understanding of the metrics that allows them to develop narrative understanding of their students' learning [80], [81]. Future work should thus examine dashboard design features that could enhance teachers' interpretation and the application of these scores as part of the formative assessment processes conducted by teachers [82]. These future directions will help make games for assessment a more integral part of the education that kids experience in school.

Acknowledgments

We wish to acknowledge the design and development team of *Shadowspect* from the Playful Journeyl Lab and Firehose, for their collaborative efforts bringing this tool to life.

References

[1] J. P. Gee, *Video Games, Learning, and "Content"*. Boston, MA: Springer US, 2009, pp. 43–53.
[2] M. Prensky, "Digital game-based learning," *Computers in Entertainment (CIE)*, vol. 1, no. 1, pp. 21–21, 2003.
[3] S. De Freitas, "Are games effective learning tools? a review of educational games," *Journal of Educational Technology & Society*, vol. 21, no. 2, pp. 74–84, 2018.
[4] R. Ibrahim and A. Jaafar, "Educational games (eg) design framework: Combination of game design, pedagogy and content modeling," in *2009 International Conference on Electrical Engineering and Informatics*, 2009, pp. 293–298.
[5] D. W. Shaffer and J. P. Gee, "The right kind of gate: Computer games and the future of assessment," *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*, pp. 211–228, 2012.
[6] F. Bellotti, R. Berta, and A. De Gloria, "Designing effective serious games: opportunities and challenges for research," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 5, no. 2010, 2010.
[7] M. Freire, Á. Serrano-Laguna, B. Manero, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón, "Game learning analytics: learning analytics for serious games," in *Learning, design, and technology*. Springer Nature Switzerland AG, 2016, pp. 1–29.
[8] M. Anderson, J. Jiang *et al.*, "Teens, social media & technology 2018," *Pew Research Center*, vol. 31, no. 2018, pp. 1673–1689, 2018.

[9] T. Fulya Eyupoglu and J. L. Nietfeld, *Intrinsic Motivation in Game-Based Learning Environments*. Springer International Publishing, 2019, pp. 85–102.

[10] Y. J. Kim, R. G. Almond, and V. J. Shute, "Applying evidence-centered design for the development of game-based assessments in physics playground," *International Journal of Testing*, vol. 16, no. 2, pp. 142–163, 2016.

[11] M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, "A systematic literature review of digital game-based assessment empirical studies: Current trends and open challenges," *arXiv preprint arXiv:2207.07369*, 2022.

[12] A. Oranje, B. Mislevy, M. I. Bauer, and G. T. Jackson, *Summative Game-Based Assessment*. Springer International Publishing, 2019, pp. 37–65.

[13] C. V. Russoniello, K. O'Brien, and J. M. Parks, "The effectiveness of casual video games in improving mood and decreasing stress," *Journal of CyberTherapy & Rehabilitation*, vol. 2, no. 1, pp. 53–66, 2009.

[14] V. J. Shute, "Stealth assessment in computer-based games to support learning," *Computer games and instruction*, vol. 55, no. 2, pp. 503–524, 2011.

[15] M. J. Gomez, J. A. Ruipérez-Valiente, P. A. Martínez, and Y. J. Kim, "Applying learning analytics to detect sequences of actions and common errors in a geometry game," *Sensors*, vol. 21, no. 4, p. 1025, 2021.

[16] Y. J. Kim and D. Ifenthaler, *Game-Based Assessment: The Past Ten Years and Moving Forward*. Springer International Publishing, 2019, pp. 3–11.

[17] R. J. Mislevy, S. Corrigan, A. Oranje, K. DiCerbo, M. I. Bauer, A. von Davier, and M. John, "Psychometrics and game-based assessment," *Technology and testing: Improving educational and psychological measurement*, pp. 23–48, 2016.

[18] K. E. DiCerbo, "Game-based assessment of persistence," *Journal of Educational Technology & Society*, vol. 17, no. 1, pp. 17–28, 2014.

[19] M. Ciman, O. Gaggi, T. M. Sgaramella, L. Nota, M. Bortoluzzi, and L. Pinello, "Serious games to support cognitive development in children with cerebral visual impairment," *Mobile Networks and Applications*, vol. 23, no. 6, pp. 1703–1714, 2018.

[20] F. Chen, Y. Cui, and M.-W. Chu, "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*, vol. 30, no. 3, pp. 481–503, 2020.

[21] R. Levy, "Dynamic bayesian network modeling of game-based diagnostic assessments," *Multivariate behavioral research*, vol. 54, no. 6, pp. 771–794, 2019.

[22] H. B. Barón, R. G. Crespo, J. P. Espada, and O. S. Martínez, "Assessment of learning in environments interactive through fuzzy cognitive maps," *Soft Computing*, vol. 19, no. 4, pp. 1037–1050, 2015.

[23] R. G. Almond, "Tips and tricks for building bayesian networks for scoring game-based assessments," in *Proceedings of the Second International Workshop on Advanced Methodologies for Bayesian Networks - Volume 9505*. Springer-Verlag, 2015, p. 250–263.

[24] S.-C. Shih, B.-C. Kuo, and S.-J. Lee, "An online game-based computational estimation assessment combining cognitive diagnostic model and strategy analysis," *Educational Psychology*, vol. 39, no. 10, pp. 1255–1277, 2019.

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

[26] S. Abdi, H. Khosravi, S. Sadiq, and D. Gasevic, "A multivariate elo-based learner model for adaptive educational systems," in *Proceedings of the 12th International Conference on Educational Data Mining*. International Educational Data Mining Society, 2019, pp. 228–233.

[27] M. Yudelson, "Elo, i love you won't you tell me your k," in *Transforming Learning with Meaningful Technologies*, M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, and J. Schneider, Eds. Springer International Publishing, 2019, pp. 213–223.

[28] S. Klinkenberg, M. Straatemeier, and H. L. van der Maas, "Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation," *Computers & Education*, vol. 57, no. 2, pp. 1813–1824, 2011.

[29] R. Pelánek, J. Papoušek, J. Řihák, V. Stanislav, and J. Nižnan, "Elo-based learner modeling for the adaptive practice of facts," *User Modeling and User-Adapted Interaction*, vol. 27, no. 1, pp. 89–118, 2017.

[30] J. L. Plass, B. D. Homer, and C. K. Kinzer, "Foundations of game-based learning," *Educational Psychologist*, vol. 50, no. 4, pp. 258–283, 2015.

[31] Y. J. Kim and J. A. Ruipérez-Valiente, "Data-driven game design: The case of difficulty in educational games," in *Addressing Global Challenges and Quality Education*, C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, and S. M. Dennerlein, Eds. Springer International Publishing, 2020, pp. 449–454.

[32] A. Denisova, C. Guckelsberger, and D. Zendle, "Challenge in digital games: Towards developing a measurement tool," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, 2017, p. 2511–2519.

[33] D. B. Clark, E. E. Tanner-Smith, and S. S. Killingsworth, "Digital games, design, and learning: A systematic review and meta-analysis," *Review of educational research*, vol. 86, no. 1, pp. 79–122, 2016.

[34] T. Greitemeyer and S. Osswald, "Effects of prosocial video games on prosocial behavior," *Journal of personality and social psychology*, vol. 98, no. 2, p. 211, 2010.

[35] L. Molyneux, K. Vasudevan, and H. Gil de Zúñiga, "Gaming social capital: Exploring civic value in multiplayer video games," *Journal of Computer-Mediated Communication*, vol. 20, no. 4, pp. 381–399, 2015.

[36] V. E. Owen and R. S. Baker, *Learning analytics for games*. MIT Press, 2020.

[37] Y. J. Kim and V. J. Shute, "The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment," *Computers & Education*, vol. 87, pp. 340–356, 2015.

[38] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Computers & education*, vol. 59, no. 2, pp. 661–686, 2012.

[39] C. Alonso-Fernández, I. Martínez-Ortiz, R. Caballero, M. Freire, and B. Fernández-Manjón, "Predicting students' knowledge after playing a serious game based on learning analytics data: A case study," *Journal of Computer Assisted Learning*, vol. 36, no. 3, pp. 350–358, 2020.

[40] J. A. Ruiperez-Valiente, M. Gaydos, L. Rosenheck, Y. J. Kim, and E. Klopfer, "Patterns of engagement in an educational massively multiplayer online game: A multidimensional view," *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 648–661, 2020.

[41] K. E. DiCerbo, "Building the evidentiary argument in game-based assessment," *Journal of Applied Testing Technology*, vol. 18, no. S1, pp. 7–18, 2017.

[42] R. J. Mislevy, R. G. Almond, and J. F. Lukas, "A brief introduction to evidence-centered design," *ETS Research Report Series*, vol. 2003, no. 1, pp. i–29, 2003.

[43] M. C. Desmarais and R. S. Baker, "A review of recent advances in learner and skill modeling in intelligent learning environments," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1, pp. 9–38, 2012.

[44] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User modeling and user-adapted interaction*, vol. 4, no. 4, pp. 253–278, 1995.

[45] J.-P. Doignon and J.-C. Falmagne, *Knowledge spaces*. Springer Science & Business Media, 2012.

[46] A. Mitrovic and B. Martin, "Evaluating adaptive problem selection," in *Adaptive Hypermedia and Adaptive Web-Based Systems*, P. M. E. De Bra and W. Nejdl, Eds. Springer Berlin Heidelberg, 2004, pp. 185–194.

[47] Y. Long and V. Aleven, "Students' understanding of their student model," in *Artificial Intelligence in Education*, G. Biswas, S. Bull, J. Kay, and A. Mitrovic, Eds. Springer Berlin Heidelberg, 2011, pp. 179–186.

[48] K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning," *Cognitive science*, vol. 36, no. 5, pp. 757–798, 2012.

[49] J. E. Beck, K.-m. Chang, J. Mostow, and A. Corbett, "Does help help? introducing the bayesian evaluation and assessment methodology," in *Intelligent Tutoring Systems*, B. P. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 383–394.

[50] A. Hershkovitz, R. S. Baker, J. Gobert, M. Wixon, and M. S. Pedro, "Discovery with models: A case study on carelessness in computer-based science inquiry," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1480–1499, 2013.

[51] S. Fancsali, "Causal discovery with models: Behavior, affect, and learning in cognitive tutor algebra," in *Proceedings of the 7th International Conference on Educational Data Mining, EDM 2014, London, UK, July 4-7, 2014*, J. C. Stamper, Z. A. Pardos, M. Mavrikis, and B. M. McLaren, Eds. International Educational Data Mining Society (IEDMS), 2014, pp. 28–35.

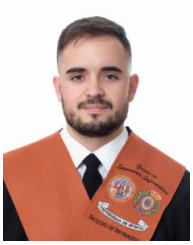
[52] M. Khajah, R. V. Lindsey, and M. Mozer, "How deep is knowledge tracing?" in *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016, Raleigh, North Carolina, USA, June 29 - July 2, 2016*, T. Barnes, M. Chi, and M. Feng, Eds. International Educational Data Mining Society (IEDMS), 2016.

[53] T. Gervet, K. Koedinger, J. Schneider, T. Mitchell *et al.*, "When is deep learning the best approach to knowledge tracing?" *Journal of Educational Data Mining*, vol. 12, no. 3, pp. 31–54, 2020.
[54] R. Scruggs, R. Baker, and B. McLaren, "Extending deep knowledge tracing: Inferring interpretable knowledge and predicting post system performance," in *Proceedings of the 28th International Conference on Computers in Education*, 2020.
[55] R. Pelánek, "Applications of the elo rating system in adaptive educational systems," *Computers & Education*, vol. 98, pp. 169–179, 2016.
[56] P. I. Pavlik, H. Cen, and K. R. Koedinger, "Performance factors analysis - A new alternative to knowledge tracing," in *Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling, Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED 2009, July 6-10, 2009, Brighton, UK*, ser. Frontiers in Artificial Intelligence and Applications, V. Dimitrova, R. Mizoguchi, B. du Boulay, and A. C. Graesser, Eds., vol. 200. IOS Press, 2009, pp. 531–538.
[57] P. I. Pavlik, L. G. Eglington, and L. M. Harrell-Williams, "Logistic knowledge tracing: A constrained framework for learner modeling," *IEEE Transactions on Learning Technologies*, vol. 14, no. 5, pp. 624–639, 2021.
[58] A. Galyardt and I. Goldin, "Move your lamp post: Recent data reflects learner knowledge better than older data." *Journal of Educational Data Mining*, vol. 7, no. 2, pp. 83–108, 2015.
[59] B. Choffin, F. Popineau, Y. Bourda, and J. Vie, "DAS3H: modeling student learning and forgetting for optimally scheduling distributed practice of skills," in *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019*, M. C. Desmarais, C. F. Lynch, A. Merceron, and R. Nkambou, Eds. International Educational Data Mining Society (IEDMS), 2019.
[60] C.-K. Yeung and D.-Y. Yeung, "Addressing two problems in deep knowledge tracing via prediction-consistent regularization," in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, ser. L@S '18. New York, NY, USA: Association for Computing Machinery, 2018.
[61] J. Zhang, R. Das, R. S. Baker, and R. Scruggs, "The cold start problem and interpretation of knowledge tracing models' predictive performance," in *Proceedings of the 14th International Conference on Educational Data Mining, EDM 2021, virtual, June 29 - July 2, 2021*, S. I. Hsiao, S. S. Sahebi, F. Bouchet, and J. Vie, Eds. International Educational Data Mining Society, 2021.
[62] C. Conati, "Intelligent tutoring systems: New challenges and directions," in *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, C. Boutilier, Ed., 2009, pp. 2–7.
[63] M. A. Sao Pedro, R. S. De Baker, J. D. Gobert, O. Montalvo, and A. Nakama, "Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill," *User Modeling and User-Adapted Interaction*, vol. 23, no. 1, pp. 1–39, 2013.
[64] Y. Cui, M.-W. Chu, and F. Chen, "Analyzing student process data in game-based assessments with bayesian knowledge tracing and dynamic bayesian networks," *Journal of Educational Data Mining*, vol. 11, no. 1, pp. 80–100, 2019.
[65] Council of Chief State School Officers, "Common core state standards initiative. high school: Geometry," http://www.corestandards.org/Math/Content/HSG/introduction/, 2021, accessed: 2022-07-30.
[66] K. R. Koedinger, P. I. P. Jr., J. C. Stamper, T. Nixon, and S. Ritter, "Avoiding problem selection thrashing with conjunctive knowledge tracing," in *Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, July 6-8, 2011*, M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. C. Stamper, Eds. www.educationaldatamining.org, 2011, pp. 91–100.
[67] C. Conati, A. Gertner, and K. Vanlehn, "Using bayesian networks to manage uncertainty in student modeling," *User modeling and user-adapted interaction*, vol. 12, no. 4, pp. 371–417, 2002.
[68] T. Murray and I. Arroyo, "Toward measuring and maintaining the zone of proximal development in adaptive instructional systems," in *Intelligent Tutoring Systems*, S. A. Cerri, G. Gouardères, and F. Paraguaçu, Eds. Springer Berlin Heidelberg, 2002, pp. 749–758.
[69] K. DiCerbo, V. Shute, and Y. J. Kim, "The future of assessment in technology rich environments: Psychometric considerations," *Learning, design, and technology: An international compendium of theory, research, practice, and policy*, pp. 1–21, 2017.
[70] P. Phobun and J. Vicheanpanya, "Adaptive intelligent tutoring systems for e-learning systems," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 2, pp. 4064–4069, 2010.
[71] K. Pliakos, S.-H. Joo, J. Y. Park, F. Cornillie, C. Vens, and W. Van den Noortgate, "Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems," *Computers & Education*, vol. 137, pp. 91–103, 2019.
[72] V. Shute, S. Rahimi, G. Smith, F. Ke, R. Almond, C.-P. Dai, R. Kuba, Z. Liu, X. Yang, and C. Sun, "Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games," *Journal of Computer Assisted Learning*, vol. 37, no. 1, pp. 127–141, 2021.
[73] S. Vanbecelaere, K. Van den Berghe, F. Cornillie, D. Sasanguie, B. Reynvoet, and F. Depaepe, "The effectiveness of adaptive versus non-adaptive learning with digital educational games," *Journal of Computer Assisted Learning*, vol. 36, no. 4, pp. 502–513, 2020.
[74] M. Seif El-Nasr and E. Kleinman, "Data-driven game development: Ethical considerations," in *International Conference on the Foundations of Digital Games*, 2020, pp. 1–10.
[75] M. P. Eladhari and E. M. Ollila, "Design for research results: experimental prototyping and play testing," *Simulation & Gaming*, vol. 43, no. 3, pp. 391–412, 2012.
[76] R. G. Almond, Y. J. Kim, G. Velasquez, and V. J. Shute, "How task features impact evidence from assessments embedded in simulations and games," *Measurement: Interdisciplinary Research & Perspectives*, vol. 12, no. 1-2, pp. 1–33, 2014.
[77] J. Huizenga, G. Ten Dam, J. Voogt, and W. Admiraal, "Teacher perceptions of the value of game-based learning in secondary education," *Computers & Education*, vol. 110, pp. 105–115, 2017.
[78] Y. S. Chee, S. Mehrotra, and J. C. Ong, "Authentic game-based learning and teachers' dilemmas in reconstructing professional practice," *Learning, Media and Technology*, vol. 40, no. 4, pp. 514–535, 2015.
[79] J. A. Ruipérez-Valiente, M. J. Gomez, P. A. Martínez, and Y. J. Kim, "Ideating and developing a visualization dashboard to support teachers using educational games in the classroom," *IEEE Access*, 2021.
[80] A. F. Wise and Y. Jung, "Teaching with analytics: Towards a situated model of instructional decision-making," *Journal of Learning Analytics*, vol. 6, no. 2, pp. 53–69, 2019.
[81] J. A. Ruipérez-Valiente and Y. J. Kim, "Effects of solo vs. collaborative play in a digital learning game on geometry: Results from a k12 experiment," *Computers & Education*, vol. 159, p. 104008, 2020.
[82] A. van Leeuwen, C. A. Knoop-van Campen, I. Molenaar, and N. Rummel, "How teacher characteristics relate to how teachers use dashboards: Results from two case studies in k-12," *Journal of Learning Analytics*, vol. 8, no. 2, pp. 6–21, 2021.

José A. Ruipérez-Valiente (SM'20) completed his B.Eng. and M.Eng. in Telecommunications at Universidad Católica de San Antonio de Murcia (UCAM) and Universidad Carlos III of Madrid (UC3M) respectively, graduating in both cases with the best academic transcript of the class. Afterwards, he completed his M.Sc. and Ph.D. in Telematics at UC3M while conducting research at Institute IMDEA Networks in the area of learning analytics and educational data mining. He completed two postdoctoral periods, one at MIT and a second one at the University of Murcia with the prestigious Spanish fellowship Juan de la Cierva. He is currently an Assistant Professor of Computer Science and Artificial Intelligence at the University of Murcia.

Yoon Jeon (YJ) Kim is an Assistant Professor of Design, Creative, and Informal Education in the Department of Curriculum and Instruction at UW–Madison. Her work centers on the topic of innovative assessment and application of playful activity design to challenge what and how we are measuring learning. YJ's playful assessment research ranges from a computer game using evidence-centered design and analytics techniques to paper-based embedded assessment tools for making. The core of her work is close collaboration with practitioners—empowering teachers to innovate around classroom assessment and use playful and authentic assessment tools that can truly impact student learning.

Ryan S. Baker is Associate Professor of Learning Sciences and Technologies at the University of Pennsylvania. Baker has developed models that can automatically detect student engagement in over a dozen online learning environments, and has led the development of the BROMP observational protocol and app for field observation of student engagement, used by over 150 researchers in 7 countries. He was the founding president of the International Educational Data Mining Society, is currently serving as Editor of the journal Computer-Based Learning in Context, is Associate Editor of the Journal of Educational Data Mining, and has co-authored published papers with over 400 colleagues.



Pedro A. Martínez received his B.Sc. in Computer Engineering at the University of Murcia, specialized in applied computing and data science. He is currently studying the MSc on New Technologies in Computer Science specialized in intelligent and knowledge-based technologies with applications in medicine. He is a member of the CyberDataLab of the University of Murcia where he is working as a research intern.

Grace C. Lin is particularly interested in measurement and playful assessments for and of learning. Her research centers around different areas of cognition and how games can be implemented to not just help people learn, but also measure elusive constructs. She received her PhD in Education from University of California, Irvine, an Ed.M. in Mind, Brain, and Education from Harvard Graduate School of Education, and a B.A. in Psychology from New York University. At UC Irvine, she was trained as a Pedagogical Fellow and conducted teaching assistant and course design PD workshops for both first year graduate students and postdocs across various disciplines. Prior to joining MIT, Grace was a postdoctoral fellow at the University of Oregon, working with nonprofit organizations and on an early childhood measures repository.