Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Modeling persistence behavior in serious games: A human-centered approach using in-game and text replays

Manuel J. Gomez *, Mariano Albaladejo-González, Félix J. García Clemente, José A. Ruipérez-Valiente

Faculty of Computer Science, University of Murcia, Calle Campus Universitario, 32, 30100, Murcia, Spain

ARTICLE INFO

Keywords:
Persistence
Serious games
Educational technology
Human-centered AI
Game-based assessment
Data annotation

ABSTRACT

Serious Games (SGs) have gained attention as powerful educational tools because of their potential to provide reliable assessments and evaluate hard-to-measure constructs and competencies that are difficult to capture using traditional forms of assessment. Specifically, this study presents a human-centered approach to model and detect persistence — a key component of successful learning outcomes — in the context of SGs. With this purpose in mind, we developed a comprehensive rubric to characterize persistence behaviors in SGs. To design the rubric, we identified a set of persistence profiles and characteristics from previous literature and elaborated a general rubric for identifying persistence behaviors at the level of individual attempts. These characteristics were then mapped onto measurable features within Shadowspect, the SG used for data collection. Following this rubric, two annotators manually labeled 1,374 level attempts from 64 students using two visualization methods: in-game and text replays. With a comprehensive dataset of 2,748 labeled attempts, we trained and evaluated Machine Learning (ML) models for each type of replay to classify persistence behaviors across four categories: *Persistence*, *Non-persistence*, *Unproductive persistence*, and *No behavior*. Our results indicate that while text-based replays enable efficient annotation with promising performance, in-game replays may provide finer detail for certain complex behaviors, highlighting the strengths and limitations of each visualization method. This work contributes the use of SGs for assessment, illustrating a transparent and adaptable AI-driven approach that enhances reliability and user-centered insights, highlighting the complementary role of human input in optimizing AI-based models to achieve meaningful, user-centered assessments in education.

1. Introduction

Technology is having a significant impact on every aspect of our lives, including gaming and education (Ullah et al., 2022). In particular, games designed for purposes other than or in addition to pure entertainment, known as Serious Games (SGs), have been increasingly studied by experts in recent years (De Gloria et al., 2014). SGs are being used across several contexts such as healthcare (Ricciardi and De Paolis, 2014), education (Zhonggen, 2019), military training (Samčović, 2018), and professional environments (Larson, 2020). They have been recognized as effective and highly engaging tools that foster skill development and ability enhancement through immersive experiences (Calvo-Morata et al., 2020; Gomez et al., 2024). In addition, SGs are being explored for their potential to provide more valid assessments compared to traditional assessment methods, such as standardized tests, multiple-choice exams, and self-report surveys (Kato and de Klerk, 2017). These conventional approaches often assess knowledge in a static manner, ignore students' thinking processes, and are not applicable to the assessment of higher-order skills (Zhu et al., 2023).

Game-Based Assessment (GBA) is a specific application of games, referring to a type of assessment that uses players' interactions as evidence to make inferences about their knowledge and skills (Gomez et al., 2022; Rafferty et al., 2015). SGs can be designed to simulate real world environments or recreate fictional contexts to prompt certain behaviors. Through these immersive and interactive environments, we can assess a broad range of skills and constructs, including competencies identified as important for success in the real world. These include non-technical skills such as communication, teamwork, and leadership (Kato and de Klerk, 2017). Moreover, GBAs are commonly used to measure 21-st century skills such as creativity, critical thinking, and persistence (Gomez et al., 2022). The importance of these skills has been widely discussed, as they help students adapt to a rapidly

* Corresponding author.
E-mail addresses: manueljesus.gomezm@um.es (M.J. Gomez), mariano.albaladejog@um.es (M. Albaladejo-González), fgarcia@um.es (F.J. García Clemente), jruiperez@um.es (J.A. Ruipérez-Valiente).

https://doi.org/10.1016/j.ijhcs.2025.103601
Received 7 November 2024; Received in revised form 12 June 2025; Accepted 8 August 2025
Available online 28 August 2025
1071-5819/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

List of Abbreviations

DT	Decision Tree
GBA	Game-Based Assessment
GBL	Game-Based Learning
HCAI	Human-Centered Artificial Intelligence
KNN	K-Nearest Neighbors
ML	Machine Learning
MLP	Multi-Layer Perceptron
RF	Random Forest
SG	Serious Game
SGD	Stochastic Gradient Descent
xAI	explainable Artificial Intelligence

evolving, technology-mediated world and equip them with lifelong learning abilities (Lazorenko and Krasnenko, 2019).

Persistence is a facet of conscientiousness that has emerged as one of the most important personality traits in predicting academic performance and several real-life outcomes (Ventura et al., 2013). Although persistence may not seem to be a uniquely 21st-century skill, it is frequently included in lists and discussions of 21st-century competencies and attributes (DiCerbo, 2014), given that modern jobs are becoming more complex and require sustained effort to complete diverse tasks (Andersson and Bergman, 2011). While the rich data generated by SGs seem promising for assessing complex skills like persistence, there are still several challenges to be addressed (Shute and Wang, 2016).

In the case of persistence, conventional metrics such as standardized tests or self-report inventories provide static snapshots that lack the granularity and real-time feedback needed for formative assessment. In contrast, log-data generated by SGs offers a vast amount of granular data, which can be utilized to identify evidence of this construct in various ways. One effective approach is to develop sophisticated models, algorithms, and metrics that enable the discovery of meaningful patterns, correlations, and insights (Luo et al., 2024) about persistence. Although past studies have explored the measurement of persistence in SGs (Ventura et al., 2013; DiCerbo, 2014), they have generally relied on rule-based methods using features or metrics inferred from in-game data. These simpler methods struggle to capture the complexity and often context-dependence nature of different persistence behaviors. For instance, determining whether a student is making an effort toward problem-solving or just engaging in blind trial-and-error attempts requires human interpretation beyond simple rules or event counting. However, building more complex models for accurate skill assessment often requires high-quality human-labeled data, which is frequently underused due to the significant cost and time investment involved (Gomez et al., 2024).

In this context, Human-Centered Artificial Intelligence (HCAI) offers a valuable perspective, emphasizing that intelligent systems should be designed with the awareness that they exist within a larger system involving human stakeholders (Riedl, 2019). By incorporating human input and feedback, HCAI promotes the development of more adaptive and personalized models, and recent research has explored methods to support the development of human-centered AI applications, particularly in designing adaptive and explainable AI systems that align with user needs (Holzinger et al., 2022). In this regard, user profiling plays a critical role by enabling more accurate and personalized inferences about user behaviors (Loh et al., 2016). This allows for the development of adaptive models that adjust to the individual learning and behavioral patterns of each user, enhancing the accuracy of skill assessments.

In this research, we focus on the skill of persistence due to its critical role in academic success and lifelong learning. We aim to model the persistence construct in SGs in a generic manner and develop a model that accurately detects users' persistence in a geometry SG through a dual approach of manual tagging and Machine Learning (ML). Specifically, we use *Shadowspect*, a 3D puzzle-based SG specifically designed to enhance geometry and spatial reasoning skills. This SG presents increasingly difficult tasks that require sustained effort and iterative problem-solving, both of which are closely aligned with the persistence construct. By adopting a human-centered approach, we combine the strengths of human expertise in annotation with the predictive power of AI models, aiming to achieve a transparent, reliable, and user-centered understanding of persistence within SGs. Specifically, we have the following objectives:

- **Persistence rubric modeling**. First, we aim to develop a detailed persistence rubric by mapping different aspects of persistence, such as decision-making, problem solving, and sustained effort, to measurable aspects using SGs data. This rubric will be the foundation for both manual annotation and ML model training.
- **Mapping persistence aspects to game features**. The second objective is to translate persistence aspects identified in the literature into specific, measurable features within *Shadowspect*, a puzzle game focused on geometry and spatial reasoning. Aligning these features with our rubric will allow us to create meaningful inputs to later support our models.
- **Manual Annotation**. This third objective involves annotating a set of replays in *Shadowspect*. We will label the same set of attempts using two types of log-data representations: in-game replays (i.e., recreating users' gameplay using the game engine) and text replays (i.e., generating a textual "pretty-printed" representation of users' actions during gameplay).
- **ML model training and evaluation**. The final objective is to train separate ML models based on both in-game and text replays. Each model will be trained independently using the annotated data from each type of replay. We will then compare the accuracy of both models, providing insights into which approach is more effective for identifying different persistence behaviors.

The rest of the paper is structured as follows: Section 2 reviews background literature on persistence measurement in SGs and the use of ML and AI for GBAs. Section 3 introduces the methodology of our research, including the persistence rubric, manual annotation procedure, and the development of ML models, and Section 4 presents the results obtained. Then, we finalize the paper with a discussion in Section 5, and conclusions and future work in Section 6.

2. Related work

In order to identify gaps and opportunities, we examined related work on skill assessment using GBAs, the use of AI in different SGs applications, and the specific approaches used to model persistence within GBA contexts. We searched relevant literature across databases such as Google Scholar, Scopus, and IEEE Xplore, using combinations of keywords including "persistence serious games", "modeling persistence behavior", "skill assessment games", and "ai serious games", among others. Next, we present an overview of the most relevant studies, which provides the foundation for our proposed approach.

2.1. Evaluating skills using GBAs

A common critique of traditional assessments is that they are ineffective at measuring knowledge, skills, and abilities beyond very simple content (Buckley et al., 2021). Game-based environments offer an engaging and interactive medium for skill development and assessment, as they are ideal for providing students with scenarios that require the application of hard-to-measure constructs and competencies, such as problem solving, persistence, and creativity, among others (Shute and Wang, 2016). These hard-to-measure constructs are difficult to assess because they often involve dynamic responses and problem-solving strategies that cannot be adequately evaluated through static testing (Yuan et al., 2015).

In this regard, Garcia et al. (2020) conducted a study on 96 Game-Based Learning (GBL) studies focused on the acquisition of “soft skills”, finding that motivation, willingness to learn, competition, and problem-solving skills were the most prominent soft skills promoted in these games. However, the review also revealed that assessment data was primarily collected through pre- and post-questionnaires. More recently, Tan and Chong (2023) conducted a similar review analyzing 24 journal articles using GBL for soft skills development. The authors found that assessment methods included questionnaires, interviews, but also the analysis of in-game data. Given the vast amount of interaction data that a SG can generate (even in a short gameplay session), the application of data mining and visualization techniques to these data can provide valuable insights regarding how players interact with the game (Freire et al., 2023), which can, in turn, be used to assess such skills.

However, GBAs in these contexts still present several challenges. People often face difficulties interpreting data collected from GBAs and building meaningful relationships from it, and researchers must first identify which variables are useful to indicate specific capabilities (Ren, 2019). Moreover, in a systematic review of 65 research papers on digital GBAs conducted by Gomez et al. (2022), 52.3% of the studies reported methodological challenges to be addressed in future research, most of which were related to the need for more complex metrics and techniques to infer new information. These challenges become even more pronounced when measuring hard-to-measure constructs and competencies, as they introduce additional difficulties, such as reliance on outdated multiple-choice and self-report measures, the lack of a clear and consensual definition of the construct or competence being assessed, and their theoretical multidimensionality, where certain dimensions may have both internal and external sources (Shute and Wang, 2016).

2.2. AI for SGs applications

AI is highly relevant to SGs applications, as it enables the creation of adaptive and personalized learning experiences and provides data-driven insights into user behaviors and learning outcomes. The analysis of large sets of user interaction data is a trend that has grown rapidly over the past few years (Freire et al., 2023), and a review of the literature on SGs shows that many previous studies have applied AI models and techniques in these contexts. Frutos-Pascual and Zapirain (2015) reviewed 129 papers, providing evidence for the development and integration of specific AI algorithms in areas such as decision-making and ML. Among the studies analyzed, the most common applications were modifying game flow and assessing or classifying users’ states and behaviors while playing (Luh et al., 2025). In addition, the education and health sectors were the most frequently addressed contexts in the reviewed studies.

For instance, Auer et al. (2022) used trace data from 621 participants to build a series of ML models to measure cognitive ability and conscientiousness scores in a GBA specifically designed to assess cognitive ability. Another example is found in Chen et al. (2020), where the authors extracted 27 behavioral features from evidence trace files to represent students’ gameplay activities. These features were then used to build a ML algorithm in order to predict students’ mastery of overall skill. Similarly, these techniques are applied in other contexts, such as military training. For example, Gombolay et al. (2017) applied various ML techniques using a SG platform developed to train Navy professionals in anti-ship missile defense tactics.

Although AI has shown significant potential in SGs for analyzing skills and behaviors, the literature still reports limitations in the complexity and generalizability of models used to infer behavioral constructs, such as persistence. Many existing AI models rely heavily on rule-based metrics, specific features, or automatically labeled data, which may limit their adaptability to other contexts and environments. This gap highlights the need for more complex but flexible AI approaches that can model and analyze complex constructs across SGs.

2.3. Modeling persistence in GBAs

Persistence is increasingly recognized as a critical factor in learning and performance, particularly in digital environments where users must address diverse challenges and adapt to complex scenarios. Previous studies have mainly focused on measuring persistence using in-game metrics or indicators, or analyzing specific patterns within the data. For example, DiCerbo (2014) assessed in-game persistence using indicators aligned with the time spent, tasks completed, and attempts after failure. Similarly, Ventura and Shute (2013) created a measure of persistence based on the time spent on unsolved and solved problems, finding that it was correlated with external measures. In contrast, Klein-Latucha and Hershkovitz (2024) analyzed persistence at a sequence level, identifying interesting patterns such as immediate re-attempts to improve scores or leaving a task before completing it, only to rerun it immediately to achieve a better score.

While these previous approaches have provided foundational insights, they largely rely on rule-based features without leveraging recent advancements in AI and ML techniques, which present new opportunities to model persistence more accurately. In contrast, our work introduces a novel hybrid approach that integrates human-centered annotation with ML to capture different persistence-related behaviors. First, we propose a generalizable rubric that maps persistence-related constructs to measurable in-game characteristics. Unlike prior work, which often uses narrow or task-specific metrics, our rubric is designed to apply across different contexts. Second, we compare the effectiveness of ML models trained on two types of replay data (in-game and text replays), both annotated using the previously developed rubric. This comparison is novel in the literature and allows us to evaluate the strengths and limitations of each replay method for modeling skills in SGs. By combining a generic rubric, human annotation, multiple data representations, and ML techniques, our research advances the state of the art and contributes a scalable and transparent framework for modeling persistence in SGs.

3. Methodology

In this Section, we present the complete methodology followed to conduct our research, which can be seen in Fig. 1. Firstly, based on previous literature on persistence in SGs, we propose a general rubric to identify different persistence behaviors. Secondly, we map the different rubric items to measurable features in our game, *Shadowspect*. In the third step, we compute the identified features based on the log-data from the game. Then, in the fourth step, we utilize our labeling tool to manually label our set of replays using two different data representations: in-game and text replays. Finally, we use the two labeled datasets to train two distinct ML models to classify persistence behaviors. Throughout this section, we explicitly reference the steps shown in Fig. 1 to help guide the reader through the process.

3.1. Context and dataset

In this study, we used data from *Shadowspect*, a 3D geometry game specifically designed for assessing math core standards and enhancing students’ visualization of relationships between 2D and 3D objects. Within the game, students can construct composite figures by combining primitive shapes such as spheres and cones, along with silhouettes, from various perspectives. Fig. 2 shows an example of a puzzle being solved. The interface displays three target silhouettes at the top, a 3D workspace for building the shape in the center, and shape manipulation

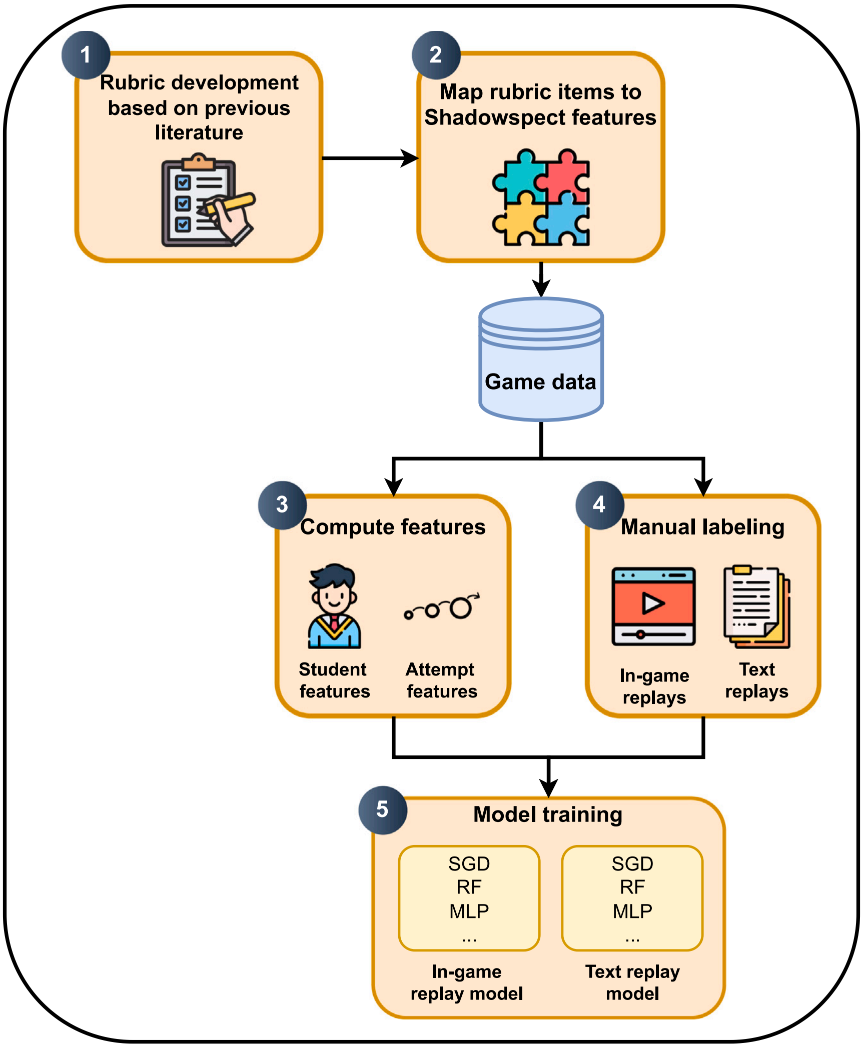


Fig. 1. Overview of the methodology to detect students' persistence.

tools at the bottom. The player is required to align the constructed figure with the given silhouettes by reasoning spatially and adjusting the position, size, orientation, and type of geometric shapes. The original version of the game in which the data was collected comprised 30 levels categorized into nine tutorial, nine intermediate, and 12 advanced levels. Firstly, tutorial levels aim to teach the basic functionality of the game, so students can learn how to build different shapes, scale and rotate them, change the camera perspective to view the structure from different angles, take snapshots of their progress, and submit their solutions to receive immediate feedback. Secondly, the intermediate levels allow students to explore more difficult puzzles without receiving so much help to solve them. Finally, the advanced levels represent a real challenge for experienced learners.

The data used in this study was collected as part of an assessment machinery that was later implemented in *Shadowspect*. For the data collection process, the team recruited seven teachers who integrated the game into their 7th grade and 10th grade math and geometry classes. Each session lasted two hours, during which the students were instructed to play for the entire session and complete as many puzzles as possible. The game provided immediate feedback on whether a puzzle attempt was successful, and the students were free to interrupt the gameplay at any time. The entire dataset comprises approximately 428,000 events performed by a total of 322 students. For this study, we selected a subset of three complete groups, totaling 64 students. This resulted in 1374 level attempts and 97,576 events, averaging 21.5 attempts and 1524 events per student. These events were recorded over a span of 67.2 h, averaging 1.05 h per student.

3.2. Rubric design

This section corresponds to step 1 in Fig. 1, which describes the design of the rubric for identifying persistence behaviors. Persistence is defined as "the tendency to remain engaged in specific goal-related activities, despite difficulties, obstacles, fatigue, prolonged frustration or low perceived feasibility" (Constantin et al., 2011). In the context of SGs, understanding and detecting persistence becomes particularly relevant. The interactive nature of SGs offers a unique environment to observe students' engagement and perseverance. In line with our research, we conducted a comprehensive analysis of the existing literature on persistence, aiming to map relevant features to measurable aspects using SGs data. This process allowed us to identify key parameters not only for annotating replays but also for constructing our predictive models.

We searched for related papers using keyword searches on indexing platforms such as *Scopus* and *Google Scholar*. To perform the search on both databases, we restricted the query to title, abstract and keywords. We included the terms "persistence", "game-based", and either "measure" or "scale". Thus, we used the following search query:

(TITLE-ABS-KEY ("persistence") AND TITLE-ABS-KEY ("game-based")) AND (TITLE-ABS-KEY ("measure") OR TITLE-ABS-KEY ("scale"))

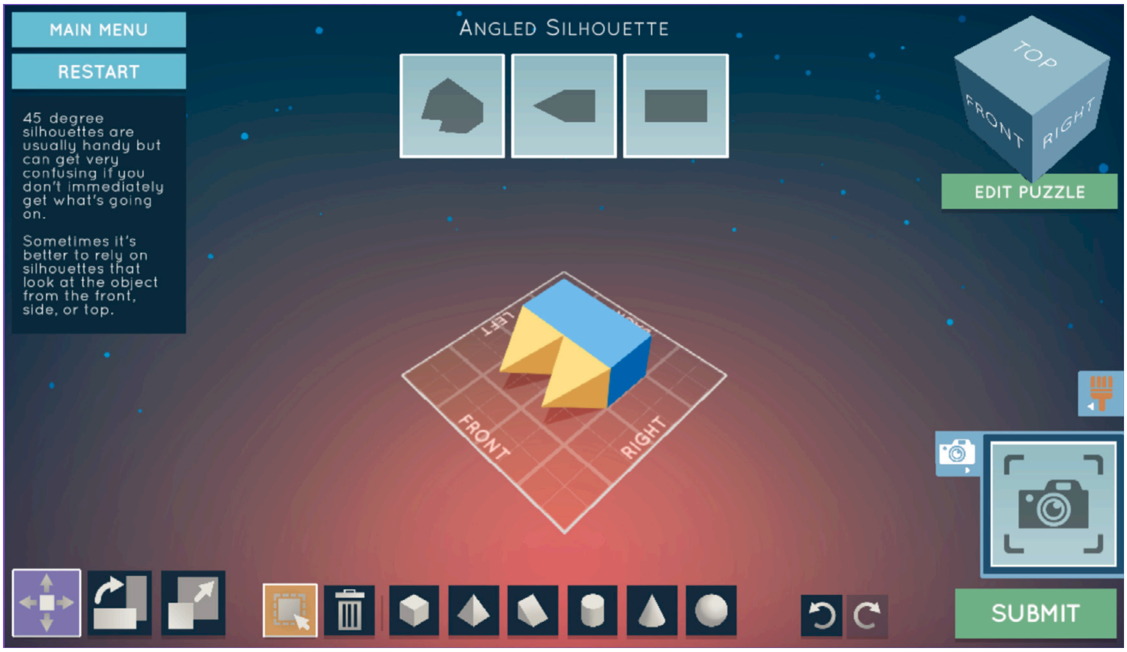

MAIN MENU
RESTART
45 degree silhouettes are usually handy but can get very confusing if you don't immediately get what's going on.
Sometimes it's better to rely on silhouettes that look at the object from the front, side, or top.
ANGLED SILHOUETTE
TOP
FRONT
RIGHT
EDIT PUZZLE
LEFT
FRONT
RIGHT
SUBMIT

Fig. 2. A puzzle example in *Shadowspect*.

To ensure a comprehensive overview of the literature, we considered a broad range of publication types, including peer-reviewed journal articles, conference papers, and book chapters. We applied no restrictions on publication date, since we wanted to include both early and more contemporary contributions to the measurement of persistence in game-based contexts. Moreover, we excluded studies based on the following exclusion criteria: (1) the term "persistence" was used in unrelated contexts (e.g., data storage, medical conditions); (2) the study was a theoretical commentary or literature review with no empirical method for assessing persistence; and (3) the research did not involve a game-based or simulated environment.

The initial selection of studies was retrieved in October 2024, generating 16 results (excluding duplicates). Additionally, we included two relevant studies from prior literature that closely aligned with our research goals. We found eight articles that were unrelated to the measurement of persistence, and three others did not present any methods for measuring persistence due to the nature of their investigation. The remaining articles explored various aspects of persistence measurement, which we describe in our results section.

Each selected article was systematically analyzed to identify persistence-related behaviors as well as any associated features or indicators. Specifically, we examined how persistence was measured or inferred, including observable in-game actions and player strategies. Through this analysis, we developed an initial set of persistence behaviors and related indicators that informed our annotation rubric. Examples of identified indicators include the total time spent solving a problem and the number of attempts or restarts during problem-solving, among others.

3.3. *Rubric-to-feature mapping*

Once we identified the different aspects of persistence from previous research, the next step (step 2 in Fig. 1) was to map these aspects to features in *Shadowspect*. Features are measurable attributes extracted from the game data that help the model identify patterns in player behavior, and they are intended to serve as inputs for our ML models. To achieve this, we first analyzed different persistence indicators and behaviors from prior studies, translating them into observable actions and decisions that a player might demonstrate in *Shadowspect*. Our goal was to design a set of features that captures:

- **General gameplay behaviors (user-level features)**: characteristics that describe a player's overall performance, such as the total time spent playing or the number of levels completed.
- **Specific actions within individual attempts (replay-level features)**: detailed metrics from each attempt, such as the time spent on the specific attempt or the number of interaction events generated.

Once these features were defined, we computed and extracted the corresponding data from *Shadowspect* and structured it into a dataset for training our predictive models. This process corresponds to step 3 of our methodology workflow (Fig. 1).

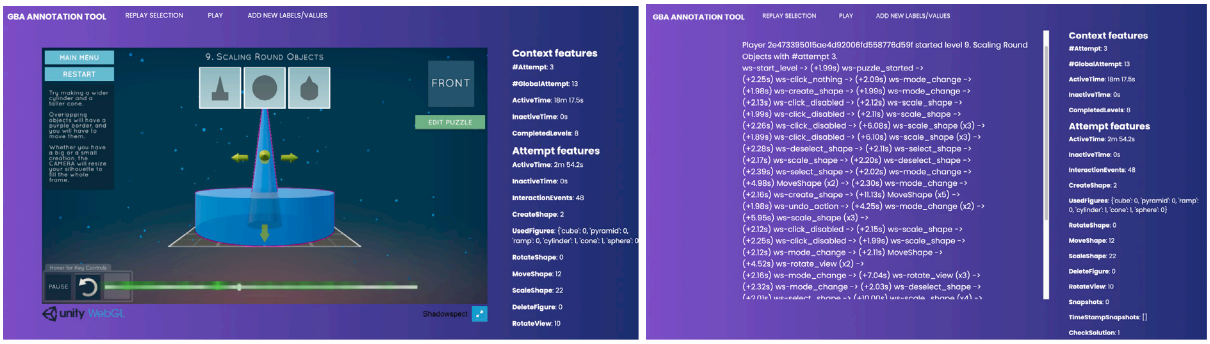
3.4. *Manual labeling*

As shown in step 4 of Fig. 1, this section describes the annotation procedure followed to label a subset of gameplay data, leveraging the rubric developed in previous steps and an existing labeling tool for GBA environments.

3.4.1. *GBA labeling tool*

In the context of GBA, manual annotation plays a crucial role for creating effective models and algorithms. To address the specific GBA research needs, previous work developed the "GBA Annotation Tool" (Gomez et al., 2024) as a Django web application. The tool supports the labeling of audio, video, and game event data, with a custom parser that integrates game event data to facilitate the analysis of gameplay performance and patterns. Users can select three annotation types and customize labels and values to meet the unique requirements of GBA scenarios. Finally, the labeled data can be exported in CSV and JSON formats for further analysis.

Fig. 3 shows two examples of replay representations using the GBA Annotation Tool. On the one hand, Fig. 3(a) shows an example of an in-game replay, which utilizes the game itself to recreate the users' gameplay and relive each action as if experiencing it in real-time. On the other hand, Fig. 3(b) shows the same replay but using a textual ("pretty-printed") representation, which includes the game's start, the number of level attempts, each action's timestamp relative to the previous action, and the final outcome (completed or not). The tool's database stores game event data and provides automatically calculated



Try making a wider cylinder and a taller cone.

Overlapping objects will have a purple border, and you will have to move them.

Whether you have a big or a small creation, the CAMERA will resize your silhouette to fill the whole frame.

Context features
#Attempt: 3
#GlobalAttempt: 13
ActiveTime: 18m 17.5s
InactiveTime: 0s
CompletedLevels: 8

Attempt features
ActiveTime: 2m 54.2s
InactiveTime: 0s
InteractionEvents: 48
CreateShape: 2
UsedFigures: {'cube': 0, 'pyramid': 0, 'ramp': 0, 'cylinder': 1, 'cone': 1, 'sphere': 0
RotateShape: 0
MoveShape: 12
ScaleShape: 22
DeleteFigure: 0
RotateView: 10

(a) In-game replay visualization.

```
Player 2e473395015ae4d92006fd558776d59f started level 9. Scaling Round
Objects with #attempt 3.
ws-start_level -> (+1.99s) ws-puzzle_started ->
(+2.25s) ws-click_nothing -> (+2.09s) ws-mode_change ->
(+1.98s) ws-create_shape -> (+1.99s) ws-mode_change ->
(+2.13s) ws-click_disabled -> (+2.12s) ws-scale_shape ->
(+1.99s) ws-click_disabled -> (+2.11s) ws-scale_shape ->
(+2.26s) ws-click_disabled -> (+6.08s) ws-scale_shape (x3) ->
(+1.89s) ws-click_disabled -> (+6.10s) ws-scale_shape (x3) ->
(+2.28s) ws-deselect_shape -> (+2.11s) ws-select_shape ->
(+2.17s) ws-scale_shape -> (+2.20s) ws-deselect_shape ->
(+2.39s) ws-select_shape -> (+2.02s) ws-mode_change ->
(+4.98s) MoveShape (x2) -> (+2.30s) ws-mode_change ->
(+2.16s) ws-create_shape -> (+11.13s) MoveShape (x5) ->
(+1.98s) ws-undo_action -> (+4.25s) ws-mode_change (x2) ->
(+5.95s) ws-scale_shape (x3) ->
(+2.12s) ws-click_disabled -> (+2.15s) ws-scale_shape ->
(+2.25s) ws-click_disabled -> (+1.99s) ws-scale_shape ->
(+2.12s) ws-mode_change -> (+2.11s) MoveShape ->
(+4.52s) ws-rotate_view (x2) ->
(+2.16s) ws-mode_change -> (+7.04s) ws-rotate_view (x3) ->
(+2.32s) ws-mode_change -> (+2.03s) ws-deselect_shape ->
(+2.01s) ws-select_shape -> (+10.00s) ws-scale_shape (x4) ->
```

Context features
#Attempt: 3
#GlobalAttempt: 13
ActiveTime: 18m 17.5s
InactiveTime: 0s
CompletedLevels: 8

Attempt features
ActiveTime: 2m 54.2s
InactiveTime: 0s
InteractionEvents: 48
CreateShape: 2
UsedFigures: {'cube': 0, 'pyramid': 0, 'ramp': 0, 'cylinder': 1, 'cone': 1, 'sphere': 0}
RotateShape: 0
MoveShape: 12
ScaleShape: 22
DeleteFigure: 0
RotateView: 10
Snapshots: 0
TimeStampSnapshots: []
CheckSolution: 1

(b) Text replay visualization.

Fig. 3. Labeling tool screenshots.

features for each replay, offering quick insights without the need to review the entire replay.

3.4.2. Annotation procedure

Throughout this study, we used the annotation tool to label game event data with in-game and text replays. In-game replays recreate users' gameplay using the game engine, enabling the annotator to visualize the users' behaviors similar to a screen recording. Text replays, on the other hand, provide a structured, "pretty-printed" representation of users' actions during gameplay. While text replays provide more limited information compared to full replays, they are likely to be very quickly to classify, and can be automatically generated using existing log files (Baker and de Carvalho, 2008). In this regard, we aimed to compare the time spent as well as the accuracy of each replay type in capturing relevant data for labeling. The annotation process was conducted at the attempt level, meaning that each individual attempt was labeled in its entirety rather than focusing on specific in-game events. There was no overlap between attempts, as each was treated as an independent unit. A total of 1,374 different attempts were labeled using both replay types, resulting in a comprehensive dataset of 2,748 labeled attempts. No time restrictions were set for the annotation process, and once completed, the labeled dataset was exported in CSV format for further analysis.

Initially, two annotators independently labeled a small subset of the dataset. Following this, the level of agreement between annotators was assessed using Cohen's kappa (κ) (DATAtab, 2025). If the resulting agreement level was low, the annotators collaboratively identified and reviewed discrepancies, resolved inconsistencies, and refined the relevant sections of the rubric accordingly. Each new iteration began with an expanded dataset incorporating additional attempts, progressively reducing inconsistencies until an acceptable inter-annotator agreement level ($\kappa > 0.7$) was achieved. Once both annotators reached satisfactory agreement, each annotator independently labeled a different portion of the data set, allowing us to efficiently scale the labeling process.

3.5. Model training and evaluation

For model training and evaluation, which corresponds to step 5 in Fig. 1, we have considered the following algorithms:

- **Adaboost**: Ensemble method that iteratively trains a sequence of weak classifiers on different subsets of the training data. After each iteration, it adjusts the weights of incorrectly classified instances to give them more emphasis in the next round.
- **Decision Tree (DT)**: Creates a tree-like model of decisions by recursively splitting data based on different feature conditions.
- **K-Nearest Neighbors (KNN)**: identifies the k closest data points to a given query point and uses their class labels (for classification) or values (for regression) to make a prediction for the query point, based on the assumption that similar points are likely to have similar outputs.
- **Multi-Layer Perceptron (MLP)**: a type of feedforward neural network where nodes are organized in multiple fully connected layers. MLPs use the backpropagation algorithm to adjust weights and enhance model accuracy.
- **Random Forest (RF)**: Ensemble method that combines the output of multiple decision trees to provide reliable predictions.
- **Stochastic Gradient Descent (SGD)**: Efficient optimization algorithm that iteratively updates the model parameters to find the minimum of the loss function using small random subsets of training data. It is convenient for large datasets, non-convex optimization problems and online learning.

To ensure reliable model evaluation, we divided the dataset into training and testing sets. Since we wanted to prevent the model from learning user-specific behaviors that could bias future predictions, we grouped all attempts from the same student into the same dataset. Thus, we decided to randomly select 70% of users as training users, and the remaining 30% as testing users. Then, the attempts corresponding to training users were included in the training dataset, and attempts made by testing users were included in the testing dataset.

For each replay type, we applied all algorithms in our pipeline, performing a grid search—an optimization technique to find the best combination of hyper-parameters in a ML algorithm. We used five-fold cross-validation, a technique that partitions the dataset into five equal subsets. The model is trained on four of these subsets and tested on the remaining one, rotating through all subsets to ensure every data point is used for both training and testing. This approach reduces bias and provides a more reliable estimate of the generalization performance of the model, helping to prevent overestimation of the true expected error. In addition, we chose balanced accuracy as our primary performance metric, which is defined as the arithmetic mean of sensitivity and specificity. It is highly useful in scenarios with imbalanced data, as it assigns equal importance to the accuracy of both the majority and minority classes (Brodersen et al., 2010). After training each model configuration separately, we selected the configurations that achieved the best average results for each replay type.

4. Results

4.1. Rubric design

Typically, literature distinguishes between *persistence* and *non-persistence*. However, recent studies (Howard and Crayne, 2019; Almeda, 2018; Fancsali et al., 2020; Kim and Miklasz, 2021) introduced a new dimension called *unproductive persistence* or *unproductive struggle*. This concept refers to situations where students continue engaging in a task despite a lack of progress, using ineffective strategies, or persisting beyond what is reasonable given the context. In our analysis, we have

Table 1
Persistence categories, selected features, and original sources.

Category	Characteristic	Description	Source(s)
Persistence	Time Investment	Total time actively solving a problem.	DiCerbo (2014), Howard and Crayne (2019), Ventura et al. (2013), Ventura and Shute (2013), Constantin et al. (2011), Shute and Wang (2016)
	Re-attempt Indicators	Number of attempts or restarts in problem solving.	DiCerbo (2014), Constantin et al. (2011), Klein-Latucha and Hershkovitz (2024), Shute and Wang (2016), Israel-Fishelson and Hershkovitz (2020)
Non-Persistence	Early Abandonment	Abandoning a problem-solving task prematurely.	Ventura et al. (2013), DiCerbo (2014), Klein-Latucha and Hershkovitz (2024)
	Inactive Solving	Lack of active engagement in problem-solving.	DiCerbo (2014)
Unproductive Persistence	Senseless Actions	Purposeless or repetitive actions in problem-solving.	Howard and Crayne (2019), Kim and Miklasz (2021)
	Repetition after Completion	Repeating problems after successfully solving them.	DiCerbo (2014), Klein-Latucha and Hershkovitz (2024), Shute and Wang (2016)

identified several characteristics associated with each of these categories. In Table 1, we present the features identified in each category, along with their definitions and corresponding sources.

Taking into account the characteristics described in Table 1, we created a rubric to systematically annotate SG data at the level of individual attempts, as shown in Fig. 4. Apart from the three categories previously identified-*persistence*, *non-persistence*, and the emerging concept of *unproductive persistence*- we also introduce the *no behavior* category, which is applicable when a replay cannot be categorized using one of the three persistence categories.

The designed rubric, which can be viewed as a multi-step decision tree, can be used to determine if an attempt performed by a player exhibits any persistence behavior. The first criterion considers whether the player has previously completed the level being labeled (repetition after completion). If so, the attempt is immediately categorized as *unproductive persistence*. For players who have not previously completed the level, the next features assessed are the time spent on problem-solving (time investment) and the number of actions performed (inactive solving). If the player spent sufficient time solving the puzzle and took a sufficient number of actions, the rubric then evaluates the problem-solving strategy (senseless actions). If the player followed a meaningful or logical strategy, the replay is categorized as *persistence*; otherwise, it is categorized as *unproductive persistence*. The concept of “meaningful or logical strategy” is inherently context-dependent, varying based on the game mechanics, level design, and task complexity. In this study, a logical strategy refers to an approach in which the player is actively trying to solve the puzzle, rather than performing random or arbitrary actions.

If the player did not spend enough time or perform enough actions, the rubric checks for level completion. If the player completed the puzzle quickly or with few actions, the replay is categorized as *no behavior*. If the level was not completed, then the rubric considers whether the player attempted to solve the puzzle or simply entered and exited without trying. If the player really tried to solve the puzzle, that means that they made an early abandonment, thus the replay is categorized as *non-persistence*. If the player entered and exited the puzzle without playing, the replay is categorized as *no behavior*.

We would like to highlight that this rubric was designed specifically to leverage human expertise and interpretation throughout the annotation process. Human annotators contribute beyond simple metrics by analyzing decision-making sequences rather than assessing individual actions or event counts. Our rubric allows annotators to identify qualitative aspects of persistence, such as distinguishing between trial-and-error approaches and strategic problem-solving, as well as recognizing specific sequences of actions that could indicate engagement or disengagement.

Table 2
User features.

Feature	Description
n_attempt	Number of attempts of the student until that moment.
n_attempt_level	Number of attempts of the student in that specific level until that moment.
active_time	Amount of active time in seconds until that moment.
inactive_time	Amount of inactive time in seconds until that moment (if the time between two events is above 30 s, the user is considered to be inactive during that period).
completed_levels	Number of unique completed levels until that moment.

Table 3
Attempt features.

Feature	Description
active_time	Amount of active time in seconds.
inactive_time	Amount of inactive time in seconds (if the time between two events is above 30 s, the user is considered to be inactive during that period).
interaction_events	Amount of interaction events generated.
check_solution	Number of times the student checked the solution in the attempt.
completed	Boolean indicating if the attempt was finally successful or not.
prev_completed	Boolean indicating if the level was previously completed by the student.

4.2. Rubric-to-feature mapping

Based on the previous characteristics, we designed a set of features crucial for detecting students’ persistence, categorized into two groups: **user features** and **attempt features**. While **user features** provide insights into the user’s overall performance and interaction patterns, **attempt features** provide specific information about what the user is doing in that specific attempt, giving us a detailed view of the actions and decisions during gameplay. Tables 2 and 3 present the user and attempt features, respectively.

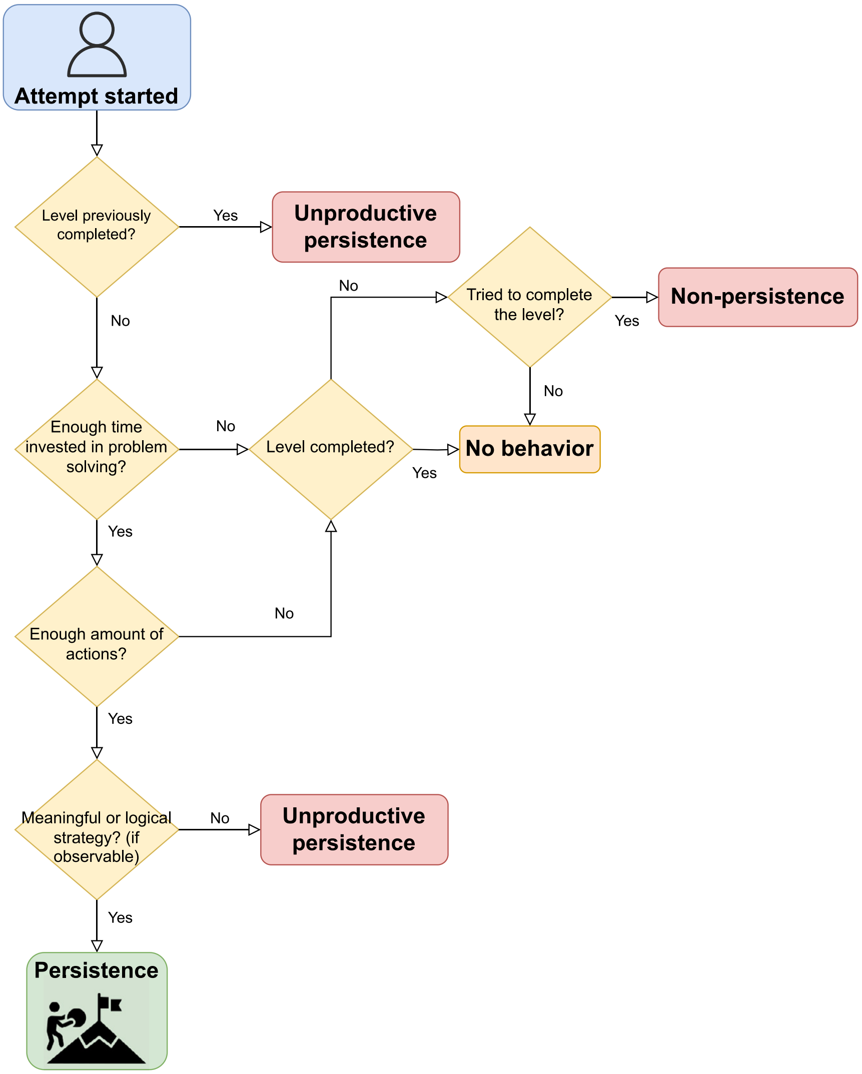

Attempt started
Level previously completed?
Yes
Unproductive persistence
No
Enough time invested in problem solving?
No
Level completed?
No
Tried to complete the level?
Yes
Non-persistence
No
No behavior
Yes
Yes
Enough amount of actions?
No
Yes
Meaningful or logical strategy? (if observable)
No
Unproductive persistence
Yes
Persistence

Fig. 4. General rubric for detecting persistence in SGs.

These features are directly aligned with the characteristics identified in the literature. Regarding *persistence*, time investment is represented through the *active_time* feature, both at the user and replay level, measuring the total time a student actively engages with problem-solving. Re-attempt indicators are captured by the *n_attempt* and *n_attempt_level* features, which count the total number of attempts a student has made overall and within specific levels, respectively. Concerning *non-persistence*, early abandonment is assessed through the combination of the *active_time* and *completed* features. If an attempt is marked as unsuccessful and with little time invested, it suggests a lack of persistence. Inactive solving is represented by the *inactive_time* feature, which measures inactivity periods during the problem-solving process. Regarding *unproductive persistence*, senseless actions are indirectly identified through the *check_solution* and *interaction_events* features, which summarize students' actions in the game. Finally, repetition after completion is reflected through the *n_attempt* and *prev_completed* features, specifically when the student continues to spend time in previously completed levels.

4.3. Manual labeling

Both annotators applied the designed rubric to label each replay, taking approximately 6.7 h for in-game replays and 3.9 h for text replays. In addition, the annotators achieved an agreement rate of 95.3% between the two different types of replay, demonstrating a high level of consistency in the annotation process. Human knowledge and expertise were crucial during the annotation process, as they allowed for context-specific decisions based on students' actions. For example, when determining whether a student performed a sufficient number of actions, it is also important to contextualize those actions. Repeating the same action 20 times or alternating between two actions ten times is not equivalent to executing a meaningful sequence of actions aimed at actually solving the proposed task. The distribution of labels among replays and categories is summarized in Table 4. It is important to note the class imbalance among the labeled categories, with *no behavior* representing the majority class in both in-game (64.9%) and text replays (66.3%).

4.4. Model training and evaluation

Following preprocessing, we configured and evaluated the ML algorithms through a five-fold cross-validation for each replay type. Since the *unproductive persistence* class has only 15 instances in in-game replays and ten in text replays, the five-fold approach ensures that this rare class is included in at least one fold per validation cycle. This is essential for imbalanced or small datasets, where higher k-values could create folds without any instances or the rare class, affecting the performance of the model on these categories.

Table 4
Count of attempts by labeled category in both types of replays.

Category	In-game replays (%)	Text replays (%)	Total (%)
No behavior	892 (64.9%)	911 (66.3%)	1803 (65.6%)
Persistence	246 (17.9%)	220 (16.0%)	466 (17.0%)
Non-persistence	221 (16.1%)	233 (17.0%)	454 (16.5%)
Unproductive persistence	15 (1.1%)	10 (0.7%)	25 (0.9%)
Total	**1374 (100%)**	**1374 (100%)**	**2748 (100%)**

Table 5
Cross-validation training results for in-game replays.

Model	Bal. Accuracy	Weighted F1	Precision
AdaBoost	0.6781	0.9138	0.9216
DT	0.7188	0.8868	0.8865
KNN	0.4198	0.6736	0.7120
MLP	0.4326	0.6502	0.6652
RF	0.7016	0.9207	0.9263
SGD	0.4216	0.5532	0.5527

Table 6
Cross-validation training results for text replays.

Model	Bal. Accuracy	Weighted F1	Precision
AdaBoost	0.6570	0.8817	0.8852
DT	0.7241	0.9094	0.9087
KNN	0.4813	0.6740	0.7049
MLP	0.4651	0.5755	0.5610
RF	0.6923	0.9129	0.9157
SGD	0.4300	0.5343	0.5458

Table 7
Test results for the in-game replays model.

Category	F1-score	Precision
No Behavior	0.94	0.96
Non-persistence	0.88	0.91
Persistence	0.70	0.65
Unproductive Persistence	0.33	0.22
Balanced accuracy	0.8	

Table 8
Test results for the text replays model.

Category	F1-score	Precision
No Behavior	0.97	0.97
Non-persistence	0.94	0.95
Persistence	0.81	0.83
Unproductive Persistence	0.18	0.12
Balanced accuracy	0.76	

The best results for each algorithm are summarized in Tables 5 and 6. Given the class imbalance in the dataset, we prioritize balanced accuracy for selecting the best results, ensuring that we give equal importance to both majority and minority classes and providing a more realistic assessment of the model's ability to generalize across all categories, particularly those with fewer instances. Results show that DT achieved the best performance in both replay types, achieving a balanced accuracy of 0.719 using in-game replays and 0.724 using text replays.

After selecting DT as the best algorithm, we assessed the generalization power of the two DT models (one for each type of replay) in the test set. The test set contained unseen data from users not employed during the models' training, configuration, and selection. Tables 7 and 8 shows the performance of both models in the test set. We report the precision and F1-score as performance metrics to indicate the models' performance in predicting each category, and the balanced accuracy serves as the final metric to evaluate the overall performance. The results indicate that the DT models performed even better during testing than in the training stage, showing a good generalization power of both models.

Although both models obtained a similar balanced accuracy (0.8 and 0.76), we can see some differences in the individual predictions of each category. For example, we observe that the text replays model achieves higher precision (0.83) when detecting the *persistence* category compared to the in-game replays model (0.65). Similarly, the text model achieves a higher precision (0.95) when predicting the *non-persistence* category. However, the in-game replays model achieves a higher precision (0.22) for predicting *unproductive persistence* labels in comparison to the text model (0.12). Finally, both models obtain a similar precision (0.96 and 0.97) when detecting the *no behavior* class.

5. Discussion

As previous literature discussed, persistence reflects learners' engagement and motivation but also plays a significant role in their overall success and skill acquisition. With GBAs providing ideal environments to assess hard-to-measure constructs, this study aimed to explore the detection of different aspects of persistence in SGs. Data generated by SGs offer low-level details of students' interactions, which can be leveraged to analyze various persistence characteristics, such as the time invested or the number of attempts made to solve a certain task. While basic metrics can provide interesting insights into students' persistence, the application of AI techniques allows for the identification of sequences and patterns that are not immediately evident. To systematically capture and categorize these aspects of persistence, we developed a general rubric designed to label individual attempts within SGs data. This rubric serves as a tool to classify persistence-related behaviors by considering specific features such as level completion or problem-solving strategies.

Using the rubric and features that were directly aligned with persistence characteristics, we were able to create two models that automatically categorize low-level student data into four persistence categories. Although the in-game replays model slightly outperformed the text model in terms of balanced accuracy, we observed that the text replays model is more precise in categorizing *no behavior*, *non-persistence* and *persistence* behaviors, while the in-game model is more precise in detecting *unproductive persistence*. Thus, although the in-game replays model demonstrates better overall performance, the text model might be a better choice when predicting persistence behaviors, since the in-game replays model only outperforms in predicting *unproductive persistence*, the minority class in our dataset. This superior performance for *unproductive persistence* may be attributed to the annotators' ability to closely examine interaction details when reviewing full replays, compared to the more limited information provided by text replays. However, using text replays significantly reduced the labeling time required. While the manual annotation process took approximately 6.7 h for labeling in-game replays, this time decreased to 3.91 h for text replays, representing a reduction of roughly 40% in labeling time. This demonstrates the efficiency of using text replays while still achieving results nearly as good as those obtained through alternative replay visualization methods.

Supervised learning relies heavily on labeled data for training models to make accurate predictions. Labeling is considered an indispensable stage of data pre-processing that can be particularly challenging

(Woodward et al., 2020). Since models learn from previous annotated data, the accuracy of these annotations directly influences the model's ability to make accurate predictions (Tu et al., 2020). In order to obtain high-quality labels, the annotators used the rubric as a structured framework to guide the annotation process, ensuring consistency and reducing subjectivity. However, human interpretation and knowledge remain essential in this process, enabling the identification of qualitative aspects of various persistence behaviors. For example, assessing whether a player demonstrates a "meaningful or logical strategy" depends on multiple contextual factors, including the specific game being played, the level's objectives, and the complexity of the task. In some cases, the optimal strategy could mean trying to iterate over a failed solution, while in others, persistence could mean methodically exploring a series of alternative solutions. Considering these contextual details is crucial to ensure that not only individual actions or event counts are considered, but also their relevance within the game's context. This highlights the importance of human judgment in distinguishing between effective problem-solving strategies and repetitive or unproductive behaviors that may superficially resemble persistence.

In addition, choosing an appropriate visualization method for the labeling process is essential, as it enables annotators to capture details that might be overlooked in simpler formats, such as context-specific actions. This richer perspective can help annotators create more accurate labels and support robust prediction capabilities. However, labels can be very costly to acquire via human labor, and researchers often tend to reuse datasets labeled from previous studies. Therefore, efficiency in data labeling is crucial, especially when dealing with large datasets. In this regard, text-based annotations provide a valuable alternative to more detailed visualization methods, as they can significantly reduce labeling time while still achieving an acceptable level of accuracy in the annotations.

Besides labels, carefully selected model features are crucial for prediction, as they capture key aspects of the data to help the model recognize patterns and make accurate predictions (Guyon and Elisseeff, 2003). For creating our models, we designed a set of features using game event data that comprised, on the one hand, a broader overview of user's interaction, and on the other hand, a more specific analysis of the current attempt being analyzed. This dual approach allows for capturing both general engagement patterns and detailed interaction details. However, none of the developed models achieved acceptable precision in classifying *unproductive persistence*, with the in-game replays model predicting only 22% of instances accurately and the text replays model achieving just 12%. This performance highlights challenges associated with identifying this particular behavior, which can be due to several factors. We note that this is the least populated class in our dataset, and the scarcity of instances makes it difficult for the model to learn meaningful associated patterns. In addition, the mapped features might not fully capture the details of this particular behavior, which indicates a need for a more comprehensive feature engineering process.

Although our results indicate lower precision in detecting *unproductive persistence*, it is important to discuss performance thresholds across different application contexts. In formative assessment environments, lower precision may still be valuable for identifying general trends and helping instructors understand students' behavior. However, within automated or adaptive learning systems designed to support teacher interventions or provide feedback, higher precision would be necessary to detect this particular behavior accurately and prevent misunderstandings. Compared to other persistence-related studies, these results align with the inherent difficulty of capturing complex cognitive and behavioral patterns solely from in-game interactions (Ventura and Shute, 2013; Shute and Wang, 2016). In addition, to mitigate potential issues arising from classification ambiguity and make predictions more transparent and interpretable, results should be presented along with a detailed report explaining the features that contributed to the model's decision. In this regard, explainable AI (xAI) techniques aim to enhance transparency and trust by providing educators and learners with insights into the rationale behind predictions, enabling them to make more informed decisions (Gomez et al., 2025). Presenting the reasoning behind the prediction, combined with a set of features specifically designed to detect this behavior, could further improve future efforts in developing a more robust model for predicting *unproductive persistence*.

Overall, we can see the great potential that SGs and assessment hold across different contexts. The integration of AI techniques for measuring complex behaviors in SGs enables a deeper understanding of learners' interactions, making it possible to capture and analyze patterns that might not be evident through basic metrics and analyses. This type of assessment provides a scalable way to evaluate skills like persistence, problem-solving, or collaboration, which are crucial both academically and professionally. We are confident that the future of games for assessment is bright, and we expect our work to contribute to the evolution of assessment practices in SGs, creating opportunities for deeper insights into student learning and engagement.

6. Conclusions and future work

This research aimed to profile and detect the persistence construct in SGs by addressing four main objectives: (1) developing a comprehensive rubric that captures key characteristics of persistence in SGs; (2) mapping identified persistence aspects to relevant game features; (3) manually labeling a set of replays from a geometry SG; and (4) building and evaluating ML models to classify different persistence behaviors based on the computed features and the labeled data. First, we identified a set of persistence profiles and characteristics from previous literature and designed a general rubric for identifying persistence behaviors at the level of individual attempts. Next, we mapped these previously identified characteristics into measurable features within the SG used for data collection (Shadowspect). Then, we manually annotated a subset of 1374 level attempts from 64 students using two visualization methods: in-game and text replays. Finally, using the dataset generated from feature mapping and manual labeling, we built and evaluated ML models to predict persistence behaviors, offering two different perspectives on how effectively these models can classify persistence in SGs. Following a human-centered approach, we highlight the value of combining human expertise with AI models to improve the robustness and precision of persistence detection models in SGs.

Although the results indicate good performance and promising applications for the developed models, this work still has some limitations. Firstly, we used a dataset from a specific SG, and relying on specific features may limit the applicability of our findings. Consequently, the rubric could require adaptations to other SGs with different mechanics or task structures. Secondly, we encountered challenges in detecting *unproductive persistence*, likely due to dataset imbalance and the small number of class instances labeled during the annotation process. Another limitation of this study is the lack of an external validation benchmark for persistence classification. As part of our future work, we plan to explore more advanced validation techniques, such as expert evaluations or comparisons with alternative classification models, to assess the reliability and precision of our approach. It is also worth noting that although two annotators were used and their agreement was verified, the methodology could be strengthened by using a larger number of annotators, with all of them labeling the whole dataset. In addition, a sample with more students from a wider range of ages and from different locations would improve the robustness of the results. Therefore, we plan to extend our dataset to include other students and classrooms. Furthermore, we aim to refine our feature engineering process to enhance the model's precision in detecting *unproductive persistence* by exploring additional interaction metrics and new features that might capture more nuanced aspects of student behavior. As part of this, visualizing trends in persistence behavior over time presents a promising direction for enhancing interpretability of player strategies. We also plan to explore alternative techniques to mitigate class imbalance and improve the classification accuracy. For instance, oversampling techniques generate synthetic data for the minority class, while undersampling reduces the dominance of majority classes, although some information is lost during the process. Another approach is class-weighted models, which assign higher importance to underrepresented classes during training. Additionally, ensemble methods, such as boosting, could enhance model robustness by improving pattern recognition in the minority class. We also intend to test xAI algorithms to improve the explainability and interpretability of the reasoning followed by AI models. Finally, taking advantage of the annotation tool used for labeling, we aim to test the developed rubric on several SGs to evaluate the effectiveness and generalizability of the proposed approach. These future directions will allow us to refine the generic rubric and enhance the robustness of our models, which will lead to more accurate classifications of persistence behaviors in a wider range of SGs.

CRediT authorship contribution statement

Manuel J. Gomez: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Mariano Albaladejo-González:** Writing – review & editing, Validation, Software, Methodology. **Félix J. García Clemente:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **José A. Ruipérez-Valiente:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Manuel J. Gomez reports financial support was provided by Fundación Séneca. José A. Ruipérez-Valiente reports financial support was provided by Fundación Séneca and Instituto Nacional de Ciberseguridad (INCIBE). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by grant 21795/FPI/22 and the REASSESS project (grant 21948/JLI/22), funded by the Séneca Foundation, Region of Murcia (Spain). This research also received support from PID2021-122466OB-I00 and PRE2022-102391, funded by MCIN/AEI/10.13039/501100011033/FEDER (Spain), and by the strategic project "Development of Professionals and Researchers in Cybersecurity, Cyberdefense and Data Science (CDL-TALENTUM)" from the Spanish National Institute of Cybersecurity (INCIBE) and by the Recovery, Transformation and Resilience Plan, Next Generation EU.

Data availability

The authors do not have permission to share data.

References

Almeda, V.Q., 2018. When Practice Does Not Make Perfect: Differentiating Between Productive and Unproductive Persistence. Columbia University.

Andersson, H., Bergman, L.R., 2011. The role of task persistence in young adolescence for successful educational and occupational attainment in middle adulthood. Dev. Psychol. 47 (4), 950.

Auer, E.M., Mersy, G., Marin, S., Blaik, J., Landers, R.N., 2022. Using machine learning to model trace behavioral data from a game-based assessment. Int. J. Sel. Assess. 30 (1), 82–102.

Baker, R., de Carvalho, A., 2008. Labeling student behavior faster and more precisely with text replays. In: Educational Data Mining 2008.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. IEEE, pp. 3121–3124.

Buckley, J., Colosimo, L., Kantar, R., McCall, M., Snow, E., 2021. Game-based assessment for education. In: OECD Digital Education Outlook 2021 Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots. OECD Publishing, p. 195.

Calvo-Morata, A., Alonso-Fernández, C., Freire, M., Martínez-Ortiz, I., Fernández-Manjón, B., 2020. Serious games to prevent and detect bullying and cyberbullying: A systematic serious games and literature review. Comput. Educ. 157, 103958.

Chen, F., Cui, Y., Chu, M.-W., 2020. Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. Int. J. Artif. Intell. Educ. 30, 481–503.

Constantin, T., Holman, A., Hojbotă, M.A., 2011. Development and validation of a motivational persistence scale. Psihologija 45 (2), 99–120.

DATAtab, 2025. Cohen's kappa – Explained simply. URL: https://datatab.net/tutorial/cohens-kappa. (Accessed: 10 March 2025).

De Gloria, A., Bellotti, F., Berta, R., 2014. Serious games for education and training. Int. J. Serious Games 1 (1).

DiCerbo, K.E., 2014. Game-based assessment of persistence. J. Educ. Technol. Soc. 17 (1), 17–28.

Fancsali, S.E., Holstein, K., Sandbothe, M., Ritter, S., McLaren, B.M., Aleven, V., 2020. Towards practical detection of unproductive struggle. In: Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21. Springer, pp. 92–97.

Freire, M., Serrano-Laguna, Á., Manero Iglesias, B., Martínez-Ortiz, I., Moreno-Ger, P., Fernández-Manjón, B., 2023. Game learning analytics: Learning analytics for serious games. In: Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy. Springer International Publishing, Cham, pp. 3475–3502.

Frutos-Pascual, M., Zapirain, B.G., 2015. Review of the use of AI techniques in serious games: Decision making and machine learning. IEEE Trans. Comput. Intell. AI Games 9 (2), 133–152.

Garcia, I., Pacheco, C., Méndez, F., Calvo-Manzano, J.A., 2020. The effects of game-based learning in the acquisition of "soft skills" on undergraduate software engineering courses: A systematic literature review. Comput. Appl. Eng. Educ. 28 (5), 1327–1354.

Gombolay, M.C., Jensen, R.E., Son, S.-H., 2017. Machine learning techniques for analyzing training behavior in serious gaming. IEEE Trans. Games 11 (2), 109–120.

Gomez, M.J., Armada Sánchez, Á., Albaladejo-González, M., García Clemente, F.J., Ruipérez-Valiente, J.A., 2025. Utilising explainable AI to enhance real-time student performance prediction in educational serious games. Expert Syst. 42 (3), e70008. http://dx.doi.org/10.1111/exsy.70008, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.70008.

Gomez, M.J., Ruipérez-Valiente, J.A., Clemente, F.J.G., 2022. A systematic literature review of game-based assessment studies: Trends and challenges. IEEE Trans. Learn. Technol. 16 (4), 500–515.

Gomez, M.J., Ruipérez-Valiente, J.A., Clemente, F.J.G., 2024. Optimizing multimedia and gameplay data labeling: A web-based tool for game-based assessment. SoftwareX 27, 101763.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (Mar), 1157–1182.

Holzinger, A., Kargl, M., Kipperer, B., Regitnig, P., Plass, M., Müller, H., 2022. Personas for artificial intelligence (AI) an open source toolbox. IEEE Access 10, 23732–23747.

Howard, M.C., Crayne, M.P., 2019. Persistence: Defining the multidimensional construct and creating a measure. Pers. Individ. Differ. 139, 77–89.

Israel-Fishelson, R., Hershkovitz, A., 2020. Shooting for the stars: Micro-persistence of students in game-based learning environments. In: Early Warning Systems and Targeted Interventions for Student Success in Online Courses. IGI Global, Hershey, PA, pp. 239–258.

Kato, P.M., de Klerk, S., 2017. Serious games for assessment: Welcome to the jungle. J. Appl. Test. Technol. 18 (S1), 1–6.

Kim, Y.J.Y., Miklasz, K., 2021. What we learned: from games to make assessment playful. In: Teaching in the Game-Based Classroom. Routledge, pp. 151–161.

Klein-Latucha, O., Hershkovitz, A., 2024. When leaving is persisting: Studying patterns of persistence in an online game-based learning environment for mathematics. J. Learn. Anal. 1–10.

Larson, K., 2020. Serious games and gamification in the corporate training environment: A literature review. TechTrends 64 (2), 319–328.

Lazorenko, L., Krasnenko, O., 2019. The importance of developing 21st century skills for advanced students. Publ. House "Baltija Publishing".

Loh, C.S., Li, I.-H., Sheng, Y., 2016. Comparison of similarity measures to differentiate players' actions and decision-making profiles in serious games analytics. Comput. Hum. Behav. 64, 562–574.

Luh, R., Eresheim, S., Tavolato, P., Petelin, T., Gmeiner, S., Holzinger, A., Schrittwieser, S., 2025. Gamifying information security: Adversarial risk exploration for IT/OT infrastructures. Comput. Secur. 151, 104287.

Luo, Y., Han, X., Zhang, C., 2024. Prediction of learning outcomes with a machine learning algorithm based on online learning behavior data in blended courses. Asia Pac. Educ. Rev. 25 (2), 267–285.

Rafferty, A.N., LaMar, M.M., Griffiths, T.L., 2015. Inferring learners' knowledge from their actions. Cogn. Sci. 39 (3), 584–618.

Ren, X., 2019. Stealth assessment embedded in game-based learning to measure soft skills: A critical review. Game-Based Assess. Revisit. 67–83.

Ricciardi, F., De Paolis, L.T., 2014. A comprehensive review of serious games in health professions. Int. J. Comput. Games Technol. 2014 (1), 787968.

Riedl, M.O., 2019. Human-centered artificial intelligence and machine learning. Hum. Behav. Emerg. Technol. 1 (1), 33–36.

Samčović, A.B., 2018. Serious games in military applications. Vojnotehnički Glasnik/Military Tech. Cour. 66 (3), 597–613.

Shute, V., Wang, L., 2016. Assessing and supporting hard-to-measure constructs in video games. In: The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications. Wiley Online Library, pp. 535–562.

Tan, B.S., Chong, K.S., 2023. Unlocking the potential of game-based learning for soft skills development: A comprehensive review. J. ICT Educ. 10 (2), 29–54.

Tu, H., Yu, Z., Menzies, T., 2020. Better data labelling with emblem (and how that impacts defect prediction). IEEE Trans. Softw. Eng. 48 (1), 278–294.

Ullah, M., Amin, S.U., Munsif, M., Safaev, U., Khan, H., Khan, S., Ullah, H., 2022. Serious games in science education. A systematic literature review. Virtual Real. Intell. Hardw. 4 (3), 189–209.

Ventura, M., Shute, V., 2013. The validity of a game-based assessment of persistence. Comput. Hum. Behav. 29 (6), 2568–2572.

Ventura, M., Shute, V., Zhao, W., 2013. The relationship between video game use and a performance-based measure of persistence. Comput. Educ. 60 (1), 52–58.

Woodward, K., Kanjo, E., Oikonomou, A., Chamberlain, A., 2020. LabelSens: enabling real-time sensor data labelling at the point of collection using an artificial intelligence-based approach. Pers. Ubiquitous Comput. 24, 709–722.

Yuan, K., Stecher, B.M., Hamilton, L.S., 2015. Feasibility of Developing a Repository of Assessments of Hard-to-measure Competencies. Rand Corporation.

Zhonggen, Y., 2019. A meta-analysis of use of serious games in education over a decade. Int. J. Comput. Games Technol. 2019 (1), 4797032.

Zhu, S., Guo, Q., Yang, H.H., 2023. Beyond the traditional: A systematic review of digital game-based assessment for students' knowledge, skills, and affections. Sustainability 15 (5), 4693.

Manuel J. Gomez is working towards a Ph.D. in Computer Science at the University of Murcia, Spain. He obtained his B.Sc. Degree with a focus on applied computing and data science, and a M.Sc. in Big Data. During this time, he completed a three-month research stay at the MIT Scheller Teacher Education Program. He is a member of the CyberDataLab in the University of Murcia, and his research interests include data mining, educational technology, game-based assessment, and natural language processing.

Mariano Albaladejo González is pursuing a Ph.D. in Computer Science at the University of Murcia, Spain. He received a B.Sc. degree in Computer Science and a M.Sc. in Big Data. His Ph.D. focuses on applying artificial intelligence to improve the training and assessment of professionals. His research interests include multimodal systems, educational technology, and reinforcement learning.

Félix J. García Clemente holds a Ph.D. in Computer Science and is a Full Professor in the area of Computer Architecture and Technology at the Faculty of Computer Science at the University of Murcia. His research activities focus on cybersecurity, cloud computing, and educational technology. As a result of his research, he has authored more than 120 publications, including journals and conference papers, and is an active member of various national and international R&D projects.

José A. Ruipérez-Valiente is an Associate Professor in the Information and Communications Engineering Department at the University of Murcia. He holds a B.Eng. and M.Eng. in Telecommunications, both earned with top class distinctions, as well as a M.Sc. and Ph.D. in Telematics from UC3M, with research conducted at Institute IMDEA Networks. His research interests span educational technology, computational social science, cybersecurity, and data science. He has co-authored over 100 scientific publications, earned over 2100 citations, and has received more than 20 awards, including the BBVA Award for Outstanding Young CS Researcher. He has contributed to 20 funded projects, mostly competitive.