Volume 12(2), 50–66. https://doi.org/10.18608/jla.2025.8809

Shaping Creativity: Exploring Creative Solutions in Serious Digital Puzzle Games

Sofia Strukova[1], Manuel J. Gomez[2] and José A. Ruipérez-Valiente[3]

Abstract

Creativity is often characterized by the capacity to generate novel ideas, explore unconventional approaches, and solve problems through intuition, curiosity, and innovative thinking. Assessing this multifaceted skill is both essential and challenging, especially in educational and game-based environments where creativity drives engagement and learning. In this study, we expanded upon existing research on creativity measurements in digital games to propose a comprehensive model for identifying creative solutions in serious digital puzzle games. This model was adapted to a geometry-focused puzzle game, with solutions manually evaluated across dimensions such as fluency, flexibility, and surprise. Using the resulting labelled dataset, we trained machine learning models to classify creative solutions, achieving promising results. These findings highlight the potential for integrated methods to advance creativity assessment and enhance educational game design, contributing to the development of tools that inspire and nurture creative thinking in learners.

Notes for Practice

- Creativity is crucial for problem-solving and innovation but challenging to assess, especially in dynamic environments like digital games. While serious games offer potential for creativity assessment, existing methods often lack integration with learning activities and fail to provide detailed, real-time feedback.
- This paper presents a rubric for identifying creative solutions in serious digital puzzle games, combining manual coding and machine learning (ML) models. It adapts a creativity assessment model to a geometry puzzle game, focusing on aspects like fluency, flexibility, and surprise, and provides an automated method for scalable creativity evaluation.
- The data-driven creativity assessment approach can be integrated into learning analytics tools to help educators monitor creative engagement, provide formative real-time feedback, and inform the design of educational games that promote creative problem-solving. Policymakers may also draw on these insights to embed creativity metrics into educational standards, reinforcing creativity as a key skill for future competencies.
- The proposed learning analytics approach, combining manual labelling and ML modelling, can serve to assess complex competencies beyond creativity. It offers a framework that could be adapted and replicated in other interactive learning environments to evaluate skills such as critical thinking, collaboration, and problem-solving.

Keywords
Creativity, game-based assessment, computational social science, data-driven evaluation, competencies.

Submitted: 15/12/2024 — **Accepted:** 16/06/2025 — **Published:** 20/08/2025

[1] *Corresponding author email: strukovas@um.es Address: Institute for Implementation Science in Health Care, University of Zürich, Universitätstrasse 84, 8006, Zürich, Switzerland. ORCID iD: https://orcid.org/0000-0002-5548-7704*
[2] *Email: manueljesus.gomezm@um.es Address: Department of Information and Communication Engineering, University of Murcia, Calle Campus Universitario 32, 30100, Murcia, Spain. ORCID iD: https://orcid.org/0000-0003-0571-2923*
[3] *Email: jruiperez@um.es Address: Department of Information and Communication Engineering, University of Murcia, Calle Campus Universitario 32, 30100, Murcia, Spain. ORCID iD: https://orcid.org/0000-0002-2304-6365*

1. Introduction

In an era characterized by rapid technological advances and the widespread adoption of digital media, the educational landscape is continuously evolving. Serious games (SGs) and game-based learning (GBL) have emerged as transformative approaches in education, offering dynamic and interactive platforms that engage students while fostering a range of cognitive and non-cognitive skills (Strukova, Ruipérez-Valiente, & Mármol, 2023). These game-based settings have demonstrated significant potential, as seen in their expanding role in global education, not only in enhancing learning outcomes and student engagement but also in providing unique opportunities for the measurement and assessment of various abilities in real time (Almeida & Simoes, 2019). Through intrinsic rewards and immediate feedback mechanisms, SGs can capture nuanced learner interactions, offering useful insights into problem-solving strategies, collaboration, and decision-making processes.

The modern world increasingly recognizes the importance of creativity. As a fundamental element of innovation, problem-solving, and adaptability, creativity is vital for thriving in a complex and uncertain future (Sternberg, 2019). It fosters innovative problem-solving and empowers individuals to approach challenges creatively, moving beyond conventional solutions and enhancing adaptability in diverse scenarios (Isaksen et al., 2023). Given creativity's significance, educational systems and workplaces are now focusing more on developing creative skills, acknowledging them as essential competencies that surpass traditional academic knowledge (Boyles, 2012). Consequently, schools must prepare future generations to identify novel challenges, devise creative solutions, adapt to the evolving ideas of others, and foster student creativity (Csikszentmihalyi & Wolfe, 2014).

The significance of creativity is universally recognized, and it has been proven that individuals who are imaginative, curious, independent thinkers and are open to new ideas, experiences, and unusual perspectives are more likely to engage in creative acts than other personality types (Erbas & Bas, 2015). However, the challenge lies in creativity's measurement and assessment. Conventional metrics, such as standardized tests and self-report inventories, offer static snapshots that lack the granularity and real-time feedback necessary for formative assessment (Plucker, 2023). Existing methods of gauging creativity often operate in isolation from other learning activities, making it challenging to integrate them into a cohesive and ongoing educational strategy.

In response to the above-mentioned challenges and due to the fact that creativity studies need both quantitative and qualitative methods supported by solid theory (Glaveanu et al., 2019), the present study aims to explore the potential of measuring creativity within the context of game-based assessment (GBA) through a dual approach of manual tagging and machine learning (ML). Recent reviews (Rafner et al., 2022) have summarized the development of digital games for creativity assessment, highlighting a range of interactive environments and evaluation methods. These works have advanced creativity assessment beyond traditional psychometric tools. However, much of the existing work focuses on open-ended creative production tasks or general problem-solving environments, while structured, domain-specific puzzle games remain less explored. Unlike previous studies that have primarily focused on measuring cognitive and content-specific skills in GBAs (Gomez et al., 2023), we introduce a novel methodology for creativity assessment. By combining the strengths of human expertise in tagging and the predictive power of ML algorithms, we seek to provide a more nuanced and actionable understanding of student creativity. Our work contributes to both the field of educational technology and the ongoing discourse about measuring complex skills by proposing an integrated approach to evaluate and foster creativity within a digitized learning environment.

Thus, after examining the state of the art regarding creativity applied to educational settings and to games, as well as application of creativity's algorithms and the existing related works, we state the following research goals (RGs):

- **RG1.** Explore creativity measurements in digital games to propose a general model for identifying creative solutions in serious digital puzzle games, with a focus on geometry and spatial reasoning.
- **RG2.** Adapt the proposed model to Shadowspect, a puzzle game on geometry and spatial reasoning, and manually code solutions on several creativity aspects.
- **RG3.** Use the labelled dataset to train ML models to classify creative solutions and evaluate their performance.

The remainder of this paper is structured as follows. In Section 2, we focus on the background of our study, uncovering the subject of creativity. In Section 3, we present our research methodology. Our findings are outlined in Section 4, while we discuss the results further in Section 5. Finally, we draw our conclusions and future research directions in Section 6.

2. Background

2.1 Creativity

The field of creativity is rooted in many branches of science, spanning across four central areas (business and economics, psychology, education and educational research, and computer science) (De-Marchis & Shchebetenko, 2022). Creativity is a broad and complex construct that is difficult to define and to quantify and is assumed to introduce new impulses into science, technology, engineering and mathematics (STEM) education.

Most widely accepted definitions of creativity highlight the importance of both originality and effectiveness, explaining that creativity involves not only generating novel and useful ideas but also refining and applying them to produce meaningful outcomes (van Broekhoven, 2023). According to Haylock (1997), creative thinking consists of four steps: (1) preparation (formulating the problem and collecting materials necessary for producing a solution), (2) incubation (all of the conscious and unconscious thinking efforts to solve problems), (3) illumination (sudden occurrence of a solution), and (4) verification (being sure that the solution is correct).

In the current age of artificial intelligence (AI), there are concerns about the need to redefine the concept of creativity. Traditional definitions of creativity formulated by scientists and artists experienced with AI still focus on the familiar elements. However, AI's expanding role in creative processes suggests a transition from focusing solely on human creativity to exploring "co-creativity," which delves into the intricate, collaborative, and dynamic interplay between humans and AI in the creative process (Wingström et al., 2022).

Building on this understanding of creativity, it is also essential to recognize the implications of educational practices in nurturing creativity among students. Research indicates that students often tend to reject original ideas in favour of more practical or common ones. By encouraging a balance between generating, evaluating, and selecting ideas, teachers can create an environment that helps students feel comfortable refining and revising their thoughts. This supportive atmosphere fosters emotional resilience in those who may be hesitant about new ideas, while encouraging those who are open to creative risks. Implementing a dual approach that integrates metacognitive and affective teaching activities can significantly enhance students' creative capacities and their willingness to explore unconventional solutions in problem-solving contexts (van Broekhoven, 2023).

2.2 Creativity Measurement

The investigation by Hernández-Torrano and Ibrayeva (2020) into 45 years of educational creativity research reveals a surge in interest, particularly over the past two decades. However, the study uncovered a critical issue: this research was predominantly conducted by a limited number of scholars. This situation underscores the urgency of incorporating diverse perspectives and methodologies (Hernández-Torrano & Ibrayeva, 2020). This necessity for diversity in approaches is particularly relevant when considering the existing methods used to measure creativity in educational settings. Walia (2019) presented a dynamic definition of creativity which has benefits for society as a whole and is best understood by considering various perspectives, namely those of the person, process, product, and press.

Several key methods are used to measure creativity, including tests of divergent thinking (which evaluate the generation of multiple possible solutions to a problem) and convergent thinking (focused on evaluating and selecting ideas). Other common measures include attitude inventories and personality and biographical inventories, as well as nominations from teachers or peers, supervisor ratings, and judgments of creative products. More direct methods involve the Consensual Assessment Technique, which uses expert judges, and self-report tools like the Creative Achievement Questionnaire (Carson et al., 2005).

Some creativity tests also assess non-cognitive creativity aspects, such as motivation (e.g., impulse control, novelty-seeking, and risk-taking) and personality traits like flexibility, independence, curiosity, and openness to new experiences. However, creativity scores are best understood as indicators of creative potential rather than actual achievements, since creative success also depends on other factors, including technical skills, field knowledge, mental health, and available opportunities (A. J. Cropley, 2000). Creativity measurement strategies can therefore extend beyond traditional tests to include metrics like grade point average (GPA) or instructor evaluations, combined with qualitative interviews and field studies to capture students' beliefs, perspectives, and behaviours, as well as the contexts that influence their creativity (Snyder et al., 2019).

A review of creativity assessments in education revealed that approximately a quarter of the analyzed articles utilized multiple assessment methods (Long et al., 2022). Half of the studies employed creative thinking or divergent thinking tests or tasks, while one-third used self-report questionnaires. These studies measured creativity across diverse domains, including science, art, and mathematics.

As digital learning environments continue to expand, new opportunities for assessing creativity through behavioural data have emerged. Within the field of learning analytics, ML techniques have been widely applied to model student performance, engagement, and problem-solving behaviours (Carroll et al., 2020; Spikol et al., 2018). However, the application of ML to specifically assess creativity remains limited, particularly in structured learning tasks such as digital puzzle games. Our work aims to bridge this gap by extending learning analytics approaches to capture creative problem-solving processes through a combination of manual labelling and computational modelling.

2.3 Related Work

Fostering creativity in education is essential for preparing students to face ambiguous problems, adapt to a rapidly changing world, and manage uncertainties. Shaheen (2010) argues that creativity supports economic growth, enhances employment, and builds national resilience, positioning schools as crucial spaces to broadly cultivate creativity, starting from early education. In addition, Collard and Looney (2014) emphasize the need for learners to be ready for unknown jobs, technologies, and challenges, highlighting adaptability as a core component of creativity. Runco (2008) expands these perspectives, asserting that creativity is not exclusive to the arts but present in various domains, with all individuals capable of thinking creatively.

Educators play a pivotal role by reinforcing and modelling creative behaviours and assigning tasks that promote original thinking, thereby preserving rather than creating creative potential.

In exploring creativity within open-ended game environments, Rahimi and colleagues (2024) examined Minecraft[1], a sandbox game that allows players to create intricate structures and designs. This study aimed to define and assess creativity in Minecraft by analyzing a sample of 52 player-built structures from videos using a modified creativity rubric. The qualitative analysis found evidence of most traditional creativity dimensions in the builds, with new dimensions, specifically realism and effort, emerging from the data. The researchers also utilized natural language processing (NLP) to automate the coding of creative elements in narrative descriptions provided by YouTube evaluators, achieving around 80% accuracy in identifying sentences related to creativity. These findings suggest that sandbox games are a conducive environment for observing and fostering creativity, with potential applications in educational settings where such games could be used to support creative expression and problem-solving skills.

In a study focused on younger learners, Hsiao and colleagues (2014) developed a digital game–based learning system (DGBL) to investigate its impact on fifth-grade students' creativity and manual skills. Over the course of a natural science unit, the study contrasted traditional instruction with the DGBL approach. Findings showed that students in the DGBL group demonstrated significant growth in creativity and engagement through flow experiences, which enabled them to perform better on manual tasks as well. This study highlights the potential of DGBL systems to nurture creativity in early education, especially in fields like science, by immersing students in hands-on problem-solving activities.

3. Methods and Materials

3.1 Methodology Overview

The entire methodology process to predict students' creativity is represented in Figure 1. First, based on related work on creativity measurements in serious digital puzzle games, we propose a general model aiming to identify creative solutions in this type of game. Accordingly, in the next step, we adapt the proposed model to a puzzle game on geometry and spatial reasoning called Shadowspect. In this context, an attempt refers to a student's individual trial to solve a specific puzzle. Respectively, we iterate through the input experiment database, identifying and storing the different types of in-game events (e.g., start a game; complete a puzzle; create, move, rotate, scale, or delete a shape) aligned with each user and the name of the puzzle in which the events occurred. Then, we identify and compute the creativity variables. Next, two independent coders manually label the video recordings in terms of surprise (the unexpectedness of the solution), fluency (the number of different ideas attempted), and flexibility (the diversity of approaches among the ideas). Finally, we use the labelled dataset to train ML models to classify creative solutions and evaluate their performance.

3.2 Exploration of Creativity Measurements (RG1)

3.2.1 Synthesis of Creativity in Games

To support the development of our general model for identifying creative solutions in serious digital puzzle games, we conducted a targeted literature review. This review synthesized how creativity is conceptualized and operationalized in prior work, with a focus on measurable aspects that could be adapted to digital gameplay environments. We looked for related papers using keyword searches on indexing platforms such as Scopus[2] and Google Scholar[3]. To perform the search on both databases, we restricted the query to title and keywords. We included the terms "creativity" and "game" and one of the terms "digital," "serious," and "educational," searching for them within the paper titles and keywords. Thus, we used the following final search query:

(TITLE(creativity) OR KEY(creativity)) AND (TITLE(digital) OR TITLE(serious) OR TITLE(educational) OR KEY(digital) OR KEY(serious) OR KEY(educational)) AND (TITLE(game) OR KEY(game))

The initial selection of studies was retrieved in November 2023, and this query generated 33 initial studies excluding duplicates. Of these, 25 studies were case-based explorations of creativity in digital games, focusing primarily on how creativity can be applied in real-life educational settings; the rest of the articles were conceptual papers explaining concrete games or surveys. Nineteen studies investigated creativity in classroom contexts. For instance, one study investigated creativity in a lab context. It found that undergraduate students who played Minecraft without specific instructions demonstrated higher post-test creativity scores than those who played with instructions, played a racing game, or watched TV (Blanco-Herrera et al., 2019).

In addition, 11 studies developed their own games for experimental purposes emphasizing creativity assessment through various approaches. Thirteen studies applied tools such as questionnaires, interviews, or structured observations to assess creativity. The rest of the articles delved into creativity aspects, which we will describe in Section 3.2.2.

[1] https://www.minecraft.net/
[2] https://scopus.com/
[3] http://scholar.google.com/

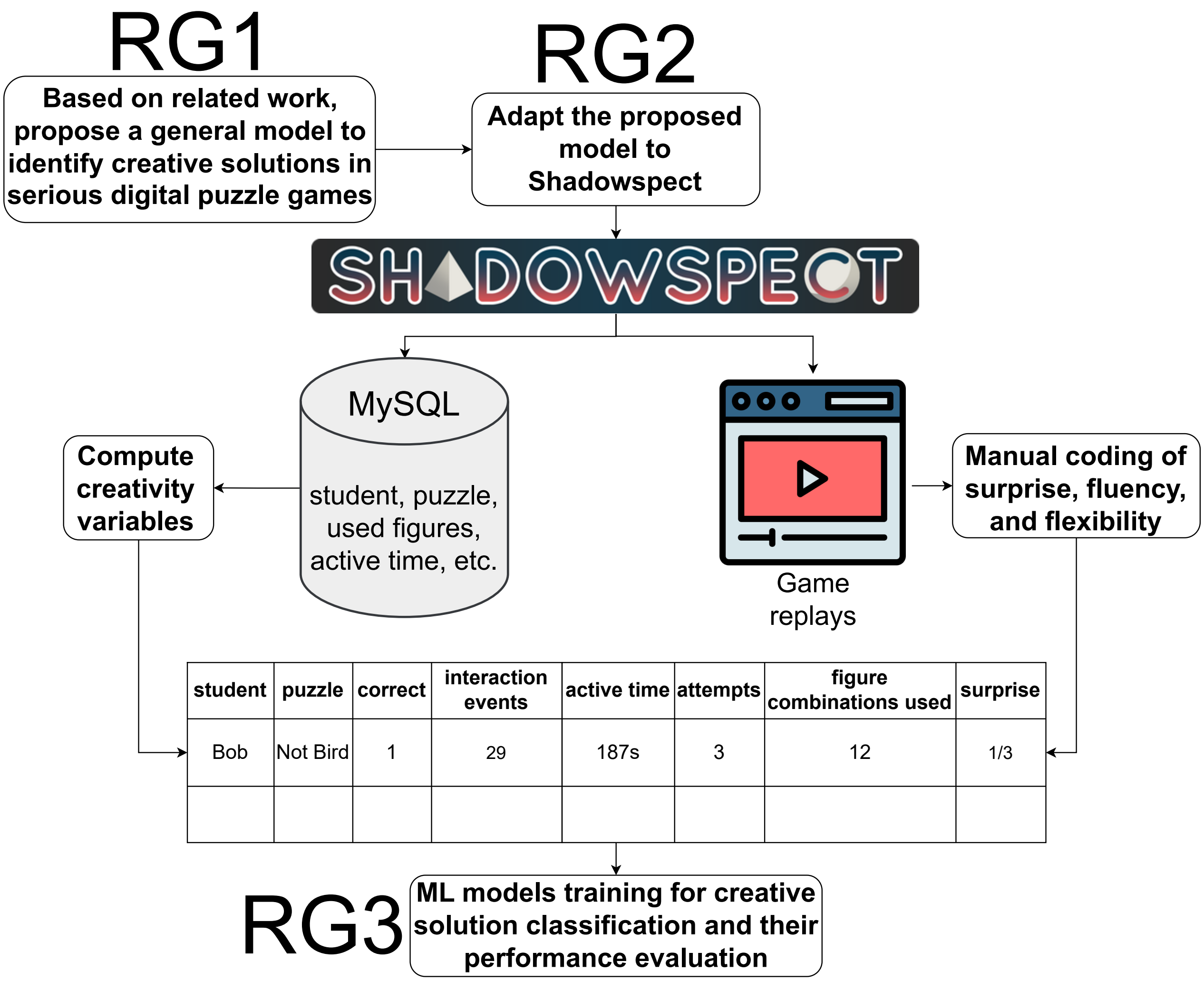

Figure 1. Overview of the methodology to predict students' creativity.

3.2.2 Creativity Aspects in Digital Games

Throughout the articles that we explored, we found the various aspects of creativity that the authors took into account, which we represent on the left side of Figure 2. One of the most common aspects across articles was originality. Seven articles used Torrance's definition of originality as the ability to produce ideas that are unusual, statistically infrequent, innovative, fresh, and not banal or obvious (Torrance, 1966). We redirected such aspects as differentness, novelty, and uniqueness to the same group. Fluency (fluidity) as the ability to produce a large number of ideas was mentioned in five articles. Four works considered flexibility as the ability to produce a large variety of ideas. Likewise, four authors observed elaboration as the ability to develop, embellish, or fill out an idea, or pay attention to details. While some aspects—such as effort, aesthetics, or realism—were mentioned less frequently, they were conceptually supported by specific studies and aligned with our game's interaction data. For instance, effort, often reflected in active engagement and time spent on a task, has been interpreted in prior work as indicative of sustained cognitive investment and motivation, which are both relevant to creative problem-solving. Based on this rationale, we included these aspects in our synthesis to ensure a comprehensive and context-sensitive model. Also, one article described content analysis, which consisted of describing the meaning of qualitative data such as texts, narrative responses, and open-ended surveys. In addition, another work evaluated generation instantiated by actions such as combine, estimate, compare, state, and planify, represented by actions such as predict, infer, hypothesize, design, and define.

Across the studies examined, authors frequently applied multiple aspects of creativity to assess outcomes within digital games. Originality, fluency, flexibility, and elaboration were some of the most frequently recurring aspects, yet authors often included additional dimensions to provide a more comprehensive view of creativity. Notably, several researchers emphasized a tripartite model, arguing that creative outputs should be not only novel and effective but also surprising (Khalil &

ISSN 1929-7750 (online).

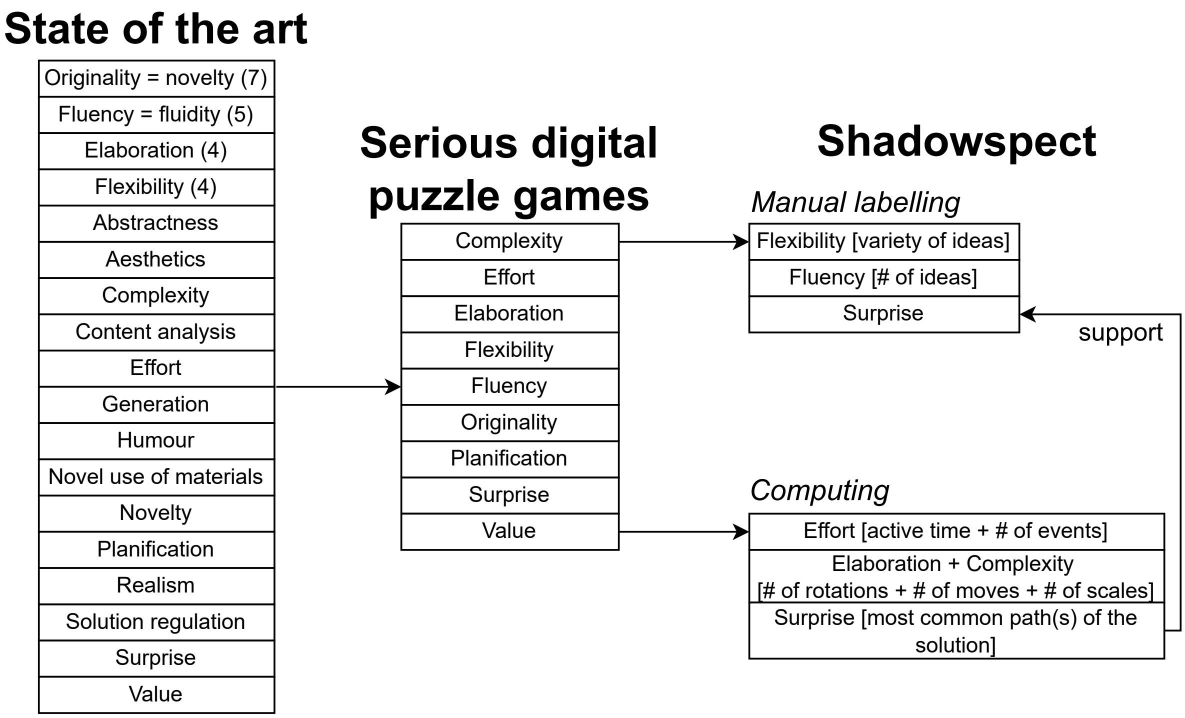

Figure 2. Overview of the creativity aspects.

Moustafa, 2022). The inclusion of surprise as a criterion highlights the expectation that creative solutions possess an element of unpredictability or cleverness that distinguishes them from random answers. Without this element, scoring based purely on statistical novelty might unintentionally favour meaningless responses, simply due to their rarity. This nuanced approach underscores the complexity of measuring creativity, especially in interactive and dynamic environments like digital games.

Rahimi and colleagues (2024) used a modified rubric for assessing a creative product in their study. They considered the following initial facets for creativity in builds of a sandbox game:

- **Surprise:** the unexpected quality of the build.
- **Elaboration:** the level of detail in the builds.
- **Aesthetics:** the quality of a build (e.g., colour, design, scale, shape, feature).
- **Realism:** the resemblance or replication of a real building or place or how the build generated a call-back or remembrance.
- **Originality:** the rareness or uniqueness of a build.
- **Effort:** the dedication and time that the user took or the difficulties that they had to overcome to create the build.
- **Complexity:** how a build was created by different elements (e.g., variations of shapes, colours, materials, items) that had an emergent quality.
- **Novel use of materials:** the use of a material or an object in an unusual way.
- **Humour:** the humorous aspect of a build.

On the other hand, other aspects not directly related to creativity were mentioned, such as domain expertise, idea evaluation skills (e.g., forecasting, appraisal, revision), feedback skills (e.g., content, timing, delivery), constraint management skills (e.g., identification, evaluation, planning), problem identification, conceptual combination, and idea generation (the ability to incorporate a previously used solution in a new way).

Based on the reviewed literature, we selected nine creativity variables that can be generally applied in serious digital puzzle games: complexity, effort, elaboration, flexibility, fluency, originality, planification, surprise, and value (presented in the centre

ISSN 1929-7750 (online).

of Figure 2). We mapped these creativity variables to measurable aspects in digital games to ensure their applicability within the context of our study. For Shadowspect, the final creativity variables are elaboration, effort, surprise, fluency, and flexibility (presented on the right side of Figure 2). Elaboration is measured by the number of solutions, moves, and scales, while effort is assessed through active time and event count. Fluency and flexibility are manually labelled, based on the number of attempts and the diversity of approaches, respectively. Surprise is also manually labelled, informed by the most common computed solution paths as a reference consisting of the used figures together with their location in the task space.

3.3 Adaptation to Shadowspect (RG2)

3.3.1 Context of the Game Environment

Figure 3. Two puzzle examples in Shadowspect.

For this article, we used data from students of 7th- and 10th-grade math and geometry classes who played Shadowspect as a part of their curriculum. Shadowspect, developed at MIT's Playful Journey Lab, is an interactive geometry puzzle game designed to engage players in solving spatial reasoning challenges through clearly defined goals, rules, and obstacles, while providing only intrinsic rewards (Ruipérez-Valiente & Kim, 2020). In the version used for this study (see Figure 3), Shadowspect includes nine tutorial levels that introduce essential gameplay mechanics, such as creating, scaling, and rotating geometric primitives. This is followed by nine intermediate levels that offer more freedom and less guidance, encouraging students to develop problem-solving skills independently. The final 12 advanced levels present increased complexity, designed to challenge students who have gained experience. Each puzzle requires students to recreate a target shape represented by silhouettes from different perspectives, using scalable and movable primitives like cubes, pyramids, ramps, cylinders, cones, and spheres. Students can adjust the camera view to examine their construction from multiple angles and use the Snapshot feature to compare their progress against the target silhouette. Upon completing a puzzle, students submit their solution for evaluation and feedback. Specifically, the system indicates which silhouettes, viewed from different perspectives, correctly match the target shapes. This automated feedback mechanism is designed to guide students in refining their solutions during subsequent attempts. Given the significance of geometry skills in STEM fields and the proven engagement and satisfaction Shadowspect offers, this game environment holds substantial potential for our investigation into creative problem-solving.

3.3.2 Puzzle Selection

To address the RGs stated earlier, we carefully considered several variables in selecting the puzzles. These included the number of students attempting each puzzle, the completion rate, average usage frequency for each shape type (cube, pyramid, ramp, cylinder, cone, sphere), average active and inactive times during attempts, number of attempts, average count of interaction events (shape creation, rotation, movement, scaling, view rotation), and whether the puzzle allowed a restricted set of shapes.

We initially excluded tutorial-level puzzles, as they offer limited scope for creativity. Additionally, puzzles with outlier characteristics—such as unusually high inactive times, low active times, fewer completed attempts, and minimal interaction events—were discarded. We prioritized puzzles with multiple possible solutions to encourage creative problem-solving.

In this way, we selected the puzzles presented in Table 1, highlighting the key metrics that guided our choices, such as student engagement, completion rates, shape usage patterns, interaction types, and time metrics. These details provide a comprehensive view of each puzzle's suitability for assessing creativity within this case study.

3.3.3 GBA Labelling Tool

In the context of GBA, manual annotation is essential for developing effective models and algorithms. To support specific GBA research needs, prior work introduced a GBA labelling tool (Gomez et al., 2024) as a Django web application. This tool enables the labelling of audio, video, and game event data through a custom parser that integrates game events to streamline the analysis of gameplay performance and patterns. Users can choose among three annotation types and define customized labels

Table 1. Puzzle details.

Puzzle Name	Unique Students	Unique Students Completed	Avg Used Cube / Pyramid / Ramp / Cylinder / Cone / Sphere	Avg Active / Inactive Time	N Attempts	Avg Interaction All Events	Avg Created / Rotated / Moved / Scaled Shapes	Avg Rotated Views	Restricted N Shapes
Angled Silhouette	128	77%	1.48 / 1.76 / 0.71 / 0.01 / 0.13 / 0.01	119s / 44s	168	33	4 / 9 / 6 / 3	8	No
Bird Fez	181	85%	1.47 / 0.99 / 0.07 / 0.63 / 0.67 / 1.09	185s / 65s	241	52	5 / 7 / 14 / 4	16	Yes
Boxes Obscure Spheres	148	70%	3.39 / 0.00 / 0.00 / 2.24 / 0.00 / 1.30	199s / 72s	244	55	7 / 4 / 16 / 2	17	Yes
More Than Meets Your Eye	83	75%	4.95 / 0.32 / 2.44 / 0.05 / 0.04 / 0.00	208s / 26s	111	61	8 / 10 / 19 / 9	11	No
Not Bird	83	69%	1.09 / 0.67 / 0.05 / 1.24 / 0.84 / 1.13	179s / 40s	126	52	5 / 7 / 10 / 6	18	Yes

and values to meet the unique requirements of GBA scenarios. Finally, the annotated data can be exported in CSV or JSON formats for further processing.

Figure 4 illustrates an example of a replay representation within the GBA labelling tool. This example shows an in-game replay, where the game itself replays the user's actions in real time, recreating the gameplay experience and allowing the annotator to review interactions in detail. The tool's database not only stores game event data but also calculates features for each replay automatically, offering quick insights without requiring a full review of the replay. Additionally, at the bottom of the figure, we can see the labelling interface, which uses selection boxes for choosing different labels and values.

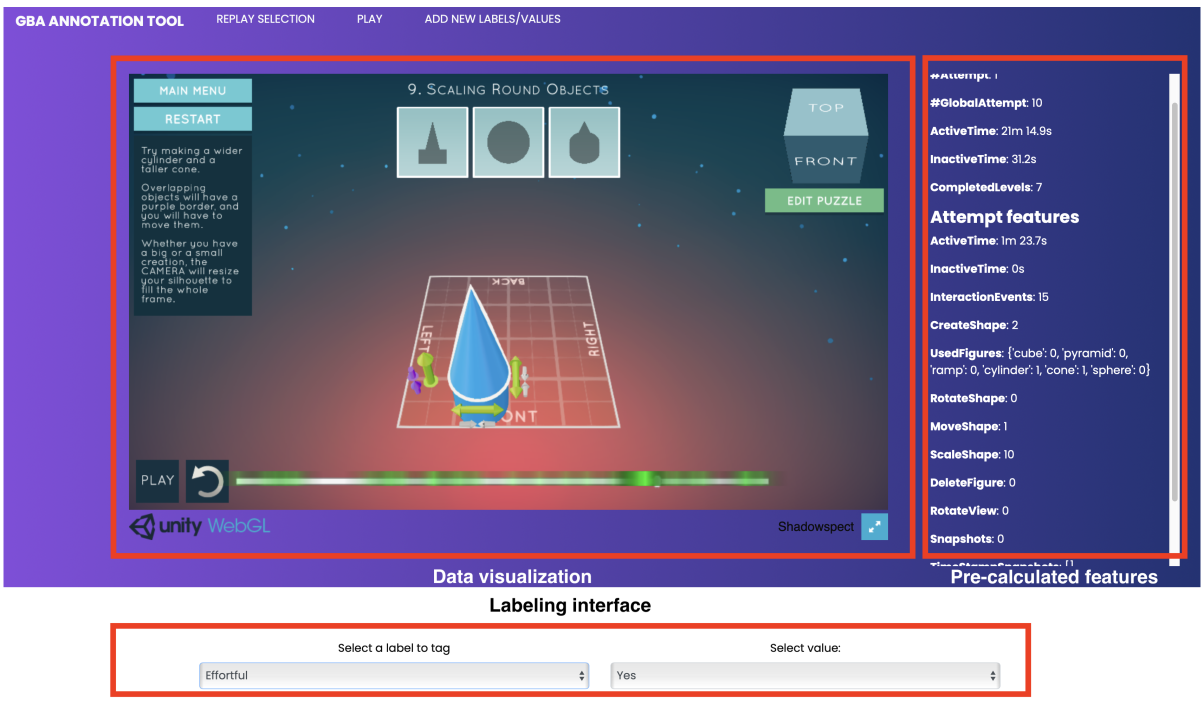

Figure 4. GBA labelling tool example.

3.3.4 Manual Coding of Creativity Aspects

The coding framework was developed by the authors and was data driven. To label each solution, the previously introduced labelling tool was used. The participants' game replays were scored regarding the genuine attempt to solve the task, fluency, flexibility, and surprise. First, two independent coders rated if the student was genuinely attempting to solve the task or was just partaking in aimless amusement. Next, they identified how many ideas were present in each of the solutions and rated how different they were on a 2-point Likert scale of "Not very different" and "Very different." Finally, the coders rated how surprising the solutions were on a 3-point Likert scale ranging from "Not surprising" to "Very surprising."

For example, in the "Not Bird" puzzle, a student might first attempt to recreate the silhouette by placing a large cylinder at the base, stacking a sphere on top, and adding a cone to represent the beak (Idea 1). Later, the student might attempt a different structure by reversing the order of the shapes or using a rectangular prism and smaller cylinder combinations instead of the original base-shape configuration (Idea 2). If the overall construction strategy between attempts changed substantially—for instance, shifting from using primarily round shapes to a combination of angular and round shapes, or altering the orientation of the assembled structure—the ideas were considered "very different." If the student's attempts varied only slightly (e.g., adjusting the cone's rotation or slightly repositioning the sphere) without changing the overall shape strategy, the ideas were considered "not very different." Surprise was manually assessed by human coders, who judged whether a solution was notably unexpected compared to typical approaches observed during the manual review. Solutions that showed an unusual choice or arrangement of shapes, or an unconventional strategy that still addressed the task requirements, were considered surprising.

The two annotators first rated 15 solutions, which were then compared. All discrepancies were discussed, aiming to discover the considerations that were used while rating, and an agreement was reached on which considerations were most important. Afterwards, the rest of the data was coded.

There were several important assumptions for the manual coding. If the solution was empty, meaning that the student did not perform any action, we assumed that the student did not genuinely attempt to solve the task. To detect a new idea, the solution should have had a new approach not necessarily starting from scratch. If there was a half solution, meaning that the student created half of the required figures, we assumed that the student did not genuinely attempt to solve the task.

The final step involved determining the ground truth labels through a process of manual coding, which integrated both the manually coded and the computed variables mentioned above (*ActiveTime*, *InteractionEvents*, *RotateShape*, *MoveShape*, and *ScaleShape*). This process required expert judgment and relied on the interplay and relative ratios of the variables, rather than adhering to strictly defined rules. While the computed variables served as support for establishing the ground truth and provided a structured foundation for the analysis, they served primarily as a decision-support tool for the coders to provide additional context rather than a definitive substitute. This approach emphasized the importance of human expertise in interpreting nuanced aspects of the data, ensuring that the coding process remained contextually informed and aligned with the study's objectives.

3.4 Model Training and Evaluation (RG3)

To evaluate creative solutions in digital puzzle games, we developed a rubric to systematically assess creativity aspects and used it to train ML models, which aim to automate the evaluation process by classifying solutions as creative based on the rubric's criteria. The process began with the preparation of the dataset, sourced from player interactions within the Shadowspect game environment excluding manually labelled aspects.

The features used to train the ML models in this study capture a wide range of interactions and performance metrics within the Shadowspect puzzle game. *GlobalAttemptID* and *SpecificPuzzleAttemptID* uniquely identify each user's game attempts, providing traceability across interactions. *LevelsCompleted* represents the progression and success in solving puzzles, serving as a measure of overall achievement. *ActiveTime* captures the duration of active engagement during gameplay, offering insights into the effort and time invested by players. *InteractionEvents* represents the total number of interactions made during gameplay, reflecting the overall activity and engagement of the player. Metrics such as *CameraRotations*, *SolutionChecks*, *RotateShape*, *MoveShape*, and *ScaleShape* capture various interaction types that contribute to problem-solving and creative exploration. Additionally, the features *CubesUsed*, *PyramidsUsed*, *RampsUsed*, *CylindersUsed*, *ConesUsed*, and *SpheresUsed* reflect the variety and frequency of geometric primitives employed by players, highlighting the complexity and creativity inherent in their solutions. Together, these features form a comprehensive dataset for assessing creativity and engagement in a GBL environment. The target output was the binary label "Creative" or "Non-creative," derived from manual coding.

We initiated the data preparation process by standardizing numerical features to ensure consistency and prevent bias in the model training. This step involved scaling continuous variables to have a mean of zero and a standard deviation of one. Categorical features were one-hot encoded to convert them into numerical representations suitable for ML algorithms. To evaluate model performance accurately, a stratified train-test split was employed, reserving 20% of the data for testing while preserving the original class distribution in both sets. These steps also served to control for the differences across different puzzles in those metrics.

Also, considering the unique characteristics of each puzzle (i.e., variations in *ActiveTime* and *InteractionEvents* as depicted in Table 1), we first decided to train a separate ML model per puzzle. However, this approach presented a challenge due to limited data for individual puzzles. Instead, we opted to incorporate puzzle identity as a dummy variable within the dataset, effectively treating puzzle differences as a categorical factor in the model. This approach allowed us to leverage all data without sacrificing the specificity of puzzle-related variations.

We evaluated four classification models: random forest (RF), support vector machine (SVM), logistic regression, and gradient boosting. We selected these models because they are well suited to relatively small, structured datasets. We did not explore deep learning approaches, as the limited sample size could lead to overfitting and would not leverage the full capabilities of such models. By including models with diverse methodologies—ensemble-based, linear, kernel-based, and gradient-based—the evaluation aimed to identify the best-performing approach for predicting creativity, ensuring robustness in assessing the dataset, and providing insights into the relative performance of different ML models. We also performed hyperparameter tuning using grid search and 10-fold cross-validation to maximize the results for each model.

The label distribution revealed that out of 220 total solutions, 53 were classified as creative, while the remaining 167 were deemed non-creative, indicating a class imbalance in the dataset. For this reason, to assess the results of the models, we used the following metrics—accuracy, precision, recall, and F1 score. Accuracy measures the overall proportion of correct predictions (both true positives and true negatives) out of all predictions. However, in imbalanced datasets, accuracy alone can be misleading, as a model might predict the majority class correctly most of the time, thus achieving high accuracy without

Table 2. Manual labelling agreement.

Puzzle solution aspect	Agreement
Real attempt to solve the task	98% (215/219)
Fluent	86% (124/144)
Flexible	89% (88/99)
Surprising	79% (113/144)

identifying the minority class effectively. Precision, which measures the proportion of true positives among all predicted positives, is crucial in this case, as we want the model to reliably identify creative solutions without falsely classifying too many non-creative ones as creative. Recall is the proportion of true positives among all actual positives, and it is important when the goal is to capture as many creative solutions as possible, even if some non-creative solutions are misclassified. Finally, the F1 score, the harmonic mean of precision and recall, provides a balanced measure that is particularly useful in the context of class imbalance, as it balances the trade-off between false positives and false negatives.

4. Results

4.1 Identification of Creative Solutions (RG1)

Figure 5 presents our foundational model for identifying creative solutions in serious digital games with open-ended tasks. This decision-making flowchart aids evaluators in determining whether a student's solution within a GBL environment can be considered creative.

The first basic part is to decide, through manual coding, whether the student is genuinely attempting to solve the task or is merely engaging in it without serious intent. If the student's intent lacks genuine problem-solving focus, the solution is immediately classified as non-creative. Second, the flowchart evaluates fluency by examining if the student proposed more than one approach to the task. This is followed by an assessment of flexibility, which checks if the ideas differ from each other in approach. For the surprise criterion, experts analyze computed variables to determine if the student avoided common solution paths. Finally, the chart assesses the effort and the elaboration aspects computationally. Effort considers whether the student put in a notable amount of work compared to peers, while elaboration assesses if the steps taken were sufficiently detailed. These aspects collectively inform the identification of creative solutions, though no single criterion is mandatory.

For manually coded variables, subjective aspects—where results may vary based on the expert's personal interpretation—should be averaged to reach the final label, ensuring a balanced assessment. Conversely, for more objective aspects, it is recommended to use the lowest value, as it provides the most certain and consistent measure.

4.2 Identification of Creative Solutions in Shadowspect (RG2)

We have adapted the model presented above to the concrete case of the Shadowspect puzzle game. These are the steps that we followed:

1. Manually code if the student was genuinely attempting to solve the task. If this is not the case, automatically consider their solution non-creative.
2. Manually code how many ideas the student proposed to solve the task.
3. Manually code if the ideas proposed by the student differ.
4. With computed common paths to solve the puzzle, consider solutions that coincide with the most common ones as non-creative. Then, for the remaining cases, manually code how surprising the solution was by paying attention to the colouring feature.
5. With the help of computed features of *ActiveTime* and *InteractionEvents*, examine if the student invested a real effort in solving the task compared to other peers.
6. With help of computed features of *RotateShape*, *MoveShape*, and *ScaleShape*, examine if the path to the solution expressed in these actions is elaborated on enough.

In Table 2, we outline the results of the manual coding for each creativity aspect. Agreement among raters was highest for assessing if a student genuinely attempted to solve the task, with 98% agreement (215 out of 219 instances). For the fluency and flexibility aspects, agreement rates were also high, at 86% (124 out of 144) and 89% (88 out of 99), respectively.

The agreement for the surprise aspect, however, was slightly lower at 79% (113 out of 144). This lower level of agreement may be due to the subjective nature of this criterion, as identifying surprising solutions often requires interpreting unique or

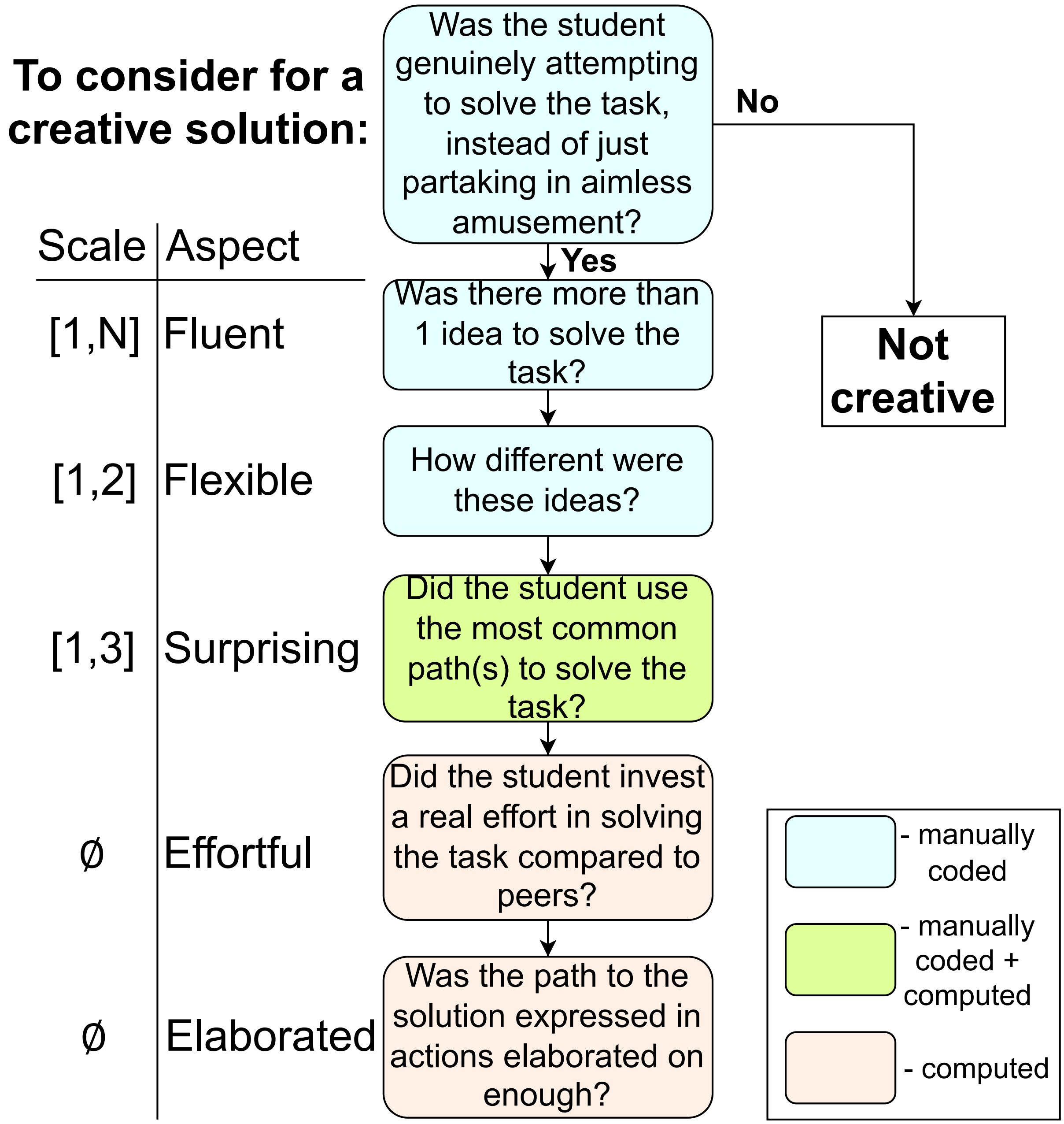

Figure 5. Proposal for a conceptual model for detecting creative solutions.

uncommon approaches. As a result, some divergence was observed in how raters judged the novelty or unexpectedness of certain solutions. Cohen's Kappa analysis confirmed a high degree of agreement between the two raters across all aspects, further supporting the reliability of the manual coding process.

Following the guidelines for manual coding presented above, genuine attempt is classified as an objective aspect, where a single "no" response from any rater classifies the solution as non-creative. Accordingly, for game replays where students did not attempt to solve the task, we did not assess fluency, flexibility, or surprise due to the absence of data. Similarly, effort and elaboration are treated as objective and use the lowest value for consistency. In contrast, surprise is considered subjective, so its final label is averaged among experts to balance individual interpretations. Moreover, if both raters agreed that a solution was not surprising, it was automatically classified as non-creative.

Table 3. Results comparison of RF, SVM, logistic regression, and gradient boosting classifier models by accuracy, precision, recall, and F1 score metrics.

	RF	SVM	Logistic regression	Gradient boosting
Accuracy	0.86	0.84	0.86	0.84
Precision	0.67	0.6	0.67	0.57
Recall	0.5	0.38	0.5	0.5
F1 score	0.57	0.46	0.57	0.53

4.3 ML Model Training and Evaluation (RG3)

4.3.1 ML Model Performance Metrics

Table 3 shows the performance metrics of four ML classification models to predict creative solutions in the Shadowspect puzzle game: RF, SVM, logistic regression, and gradient boosting. These results were obtained from an independent test dataset, representing 20% of the original data. Both RF and logistic regression demonstrated the highest accuracy at 0.86, indicating that approximately 86% of the solutions were correctly classified as creative or non-creative. The SVM and gradient boosting models showed slightly lower accuracy at 0.84. In terms of precision, RF and logistic regression achieved the best performance with 0.67, meaning that when these models predicted a solution as creative, they were correct 67% of the time. The SVM model showed a lower precision of 0.6, while gradient boosting had the lowest precision at 0.57. The recall metrics revealed more variability. RF, logistic regression, and gradient boosting all achieved a recall of 0.5, suggesting that they identified 50% of the actual creative solutions in the test dataset. The SVM model had a notably lower recall of 0.38, indicating that it was less effective at identifying creative solutions. The F1 scores, which provide a balanced measure of precision and recall, showed RF and logistic regression performing equally at 0.57. Gradient boosting scored 0.53, and the SVM model had the lowest F1 score of 0.46.

These results suggest that RF and logistic regression are the most promising models for predicting creative solutions in this context. The consistent performance across these two models indicates robustness in the approach to identifying creativity within the Shadowspect puzzle game environment.

4.3.2 Exploratory Evaluation of the ML Model

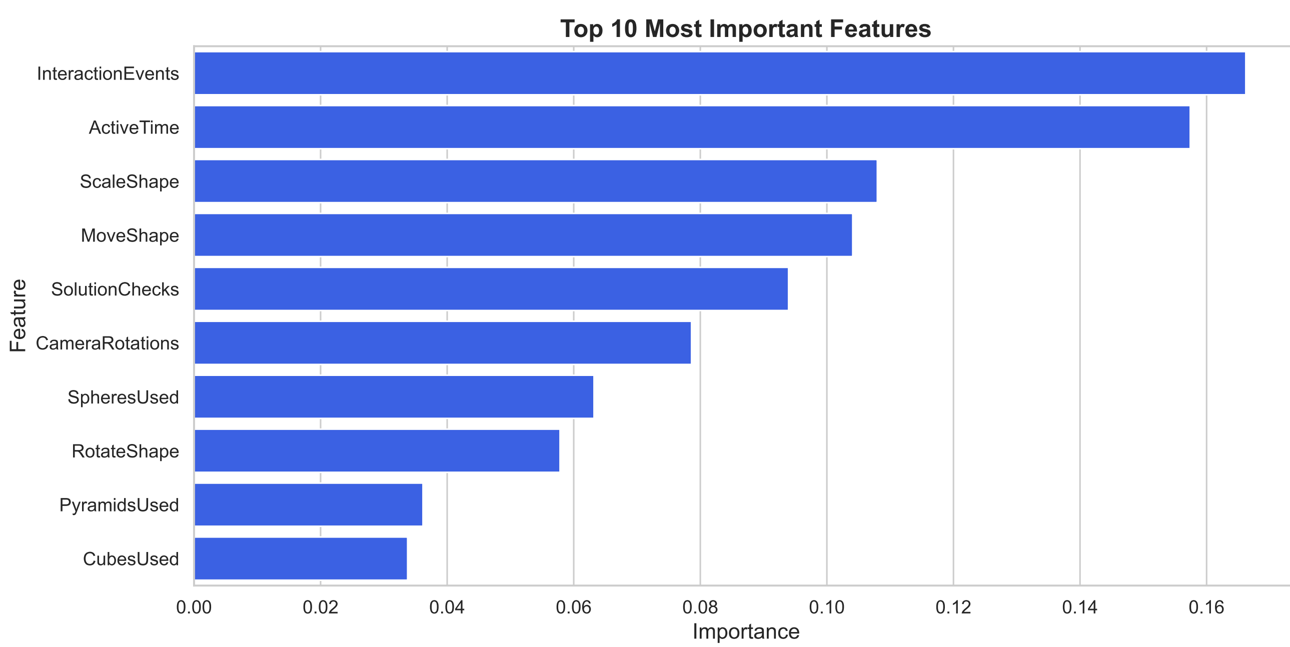

Figure 6. Feature importances.

Figure 6 illustrates the 10 most important features in predicting creativity within the Shadowspect game. These features were derived from the RF model, which showed the best results. Feature importance was calculated using a special attribute provided by the RF implementation in scikit-learn, which measures the mean and standard deviation of the reduction in impurity brought by each feature across all trees in the ensemble. These features indicate that creative solutions are characterized by active engagement, diverse shape manipulations, exploration of different perspectives, and the use of different shapes. Interaction events, active time, scaling, moving, and checking solutions are strongly correlated with creativity.

In addition, we explored the relationship between creativity and puzzle completion. Among non-creative solutions, the majority (134 out of 167) were not completed, while only 33 were completed. Conversely, for creative solutions, a smaller but still notable portion (48 out of 52) were not completed, with only four being successfully completed.

Finally, we performed another experiment to analyze the proportion of creative solutions among the total solutions attempted for six different puzzles within the Shadowspect game environment. The puzzle "Boxes Obscure Spheres" had the highest total number of solutions, with 74 submissions, of which 18 were classified as creative, resulting in a creative solution rate of 24.32%. The puzzle "Not Bird" exhibited a smaller sample size, with 27 solutions submitted, nine of which were creative, yielding the highest percentage of creative solutions at 33.33%. Similarly, the puzzle "Unnecessary" also achieved a creative solution percentage of 33.33%, though with fewer total attempts (21 solutions, seven creative). For the puzzle "Angled Silhouette," 32 total solutions were recorded, eight of which were creative, corresponding to a creative solution percentage of 25%. A comparable percentage was observed for "More Than Meets Your Eye," which had 24 total solutions and six creative ones. In contrast, the puzzle "Bird Fez" demonstrated the lowest percentage of creative solutions, with only four out of 41 solutions classified as creative, resulting in a rate of 9.76%.

5. Discussion

5.1 Results Interpretation

The results presented above reveal that RF and logistic regression performed best, likely due to their ability to effectively capture patterns in the data without overcomplicating the classification process. Their higher precision suggests that they are particularly effective at distinguishing creative solutions from non-creative ones when clear indicators are present. The slightly lower performance of SVM and gradient boosting could be related to the nature of the data; creative solutions might exhibit subtle, complex patterns that are harder to generalize. Additionally, the relatively modest recall across all models indicates that creative solutions are more challenging to identify consistently, possibly explained by the inherent subjectivity in defining creativity. These results suggest that while the models are effective in some aspects, the complexity and nuanced nature of creativity require careful consideration of both the data and the algorithms used to predict it. Further exploration of additional features or context-specific data could improve the models' ability to capture these subtleties.

Notably, there is a scarcity of comparable studies that have attempted to analyze creativity in serious digital puzzle games through a similar dual approach of manual labelling and ML. Existing research often relies only on manual annotation, computational analysis of performance metrics, or survey-based techniques, but not a combination of several of these. Furthermore, measuring creative problem-solving is particularly challenging due to the complexity of assessing individual expertise and the ambiguity surrounding artistic abilities (Strukova, Marco, et al., 2023). However, another study analyzed final creative products using content analysis and a predefined rubric (Rahimi et al., 2024). Unlike our method, which integrates detailed analyses of in-game behaviours with manual labelling and ML to identify creative solutions, this study focused primarily on evaluating the final artifact and its overall outcome. Furthermore, while Ellison and Drew (2019) explored Minecraft as a tool for enhancing boys' creativity in writing, their research primarily used a pre- and post-test writing assessment alongside focus group interviews. They assessed writing samples using a criterion-referenced analysis tool focused on vocabulary and sentence structure, and they gathered qualitative data from students' perceptions. While they found some evidence of improved creativity in writing and positive student perceptions, their study did not employ the combination of computational methods and manual coding for directly assessing creativity in the gameplay process that our study has introduced. In this way, our approach represents a significant novelty that shows a shift from outcome-based assessments, underscoring the importance of examining the process of creative solution-finding in GBL environments.

Our additional experiments showed that non-creative solutions are overwhelmingly associated with incomplete tasks, suggesting that the lack of creative engagement may correspond to reduced problem-solving effectiveness or persistence. Moreover, while creative solutions are less frequently observed overall, they also face high rates of non-completion. This may imply that creative problem-solving attempts are more challenging and less likely to result in completion under the current task conditions. From these trends, it is evident that fostering creativity alone is not sufficient for ensuring task success. Supporting structures—such as guidance, feedback, or resources—might be necessary to bridge the gap between creative approaches and successful outcomes. These findings underscore the importance of a balanced focus on fostering both creative exploration and effective task completion strategies in educational GBL environments.

While a correlation between the number of attempts and the creativity of solutions was not evident, it is important to note that creativity can manifest at various stages of the puzzle-solving process. This suggests that creative thinking is not solely dependent on trial and error, but rather on a complex interplay of cognitive processes, including divergent thinking, original thought, and risk-taking. Further research is needed to explore the factors that influence creative problem-solving within GBAs, such as the design of the game itself, the level of challenge, and the individual characteristics of the learners.

Lastly, the results which represent the proportion of creative solutions among the total solutions highlight variability in creativity across puzzles, suggesting that certain puzzle designs may better facilitate or encourage creative problem-solving than others. The comparatively higher rates of creative solutions in "Not Bird" and "Unnecessary" imply that these puzzles may offer more opportunities for diverse and original approaches. Conversely, the low creative engagement in "Bird Fez" suggests that its structure may be more restrictive or less conducive to creative exploration. These findings can inform the design of future puzzles aimed at fostering creativity within GBL environments.

It is important to mention that the game was not explicitly designed to foster creativity and that it includes puzzles with varying difficulty levels, which may influence the capacity for creative problem-solving. Easier puzzles may provide limited opportunities for creativity, as their solutions are often straightforward and constrained by design. Conversely, more challenging puzzles like "Unnecessary" may encourage players to explore diverse strategies, leading to higher creative engagement. These findings underscore the importance of puzzle design in fostering creativity, even in environments where creativity is not the primary focus.

5.2 Implications

Our study demonstrates broad applicability to serious digital puzzle games, emphasizing their potential as educational tools. By leveraging such games, educators can integrate creative problem-solving into the curriculum in a way that aligns with both learning and assessment goals. The focus on open-ended tasks enables students to engage with problems requiring creative thinking, aligning with our RGs, which aim to evaluate creativity and its connection to task success. Specifically, the findings suggest that fostering creativity through digital games supports the exploration of multiple solutions and innovative approaches, which can translate into improved spatial reasoning and problem-solving skills.

Promoting creativity in mathematics classrooms is essential, as it encourages students to move beyond rote memorization and adopt a more exploratory approach to learning. For instance, creativity can be cultivated by providing open-ended problems that allow students to think divergently (Haylock, 1997), approach problems in various ways, and develop multiple solutions (Leikin, 2009). Our findings suggest that this aligns well with digital games like Shadowspect, where tasks inherently support these practices by requiring students to construct, modify, and test geometric shapes in novel ways. Encouraging students to generate their own problems or hypotheses and test their validity (Silver, 1997) within such environments could further enhance their creative and critical thinking skills.

In the context of fostering creativity, teachers play a critical role. Cropley's framework highlights the importance of creating conditions that support independence, integration, and flexibility in learning (A. Cropley, 1997), which are directly relevant to the GBL environments explored in our study. For instance, Shadowspect tasks encourage independence by requiring students to experiment with solutions, while the interactive and flexible gameplay design supports integration of diverse approaches and ideas. Teachers can extend these opportunities by guiding students to reflect on their processes and evaluate their learning outcomes, reinforcing the iterative nature of creativity.

Furthermore, digital puzzle games can help educators delay judgment on student solutions, allowing learners to explore unconventional ideas without immediate pressure. This aligns with fostering a classroom culture where questions are encouraged and failure is reframed as an opportunity for growth—a principle supported by our study's observation that creative attempts, even when incomplete, contribute to skill development.

Finally, educators should recognize the role of frustration as a learning tool. In our study, students frequently encountered challenges when pursuing creative solutions. Teachers can harness such moments to teach resilience and problem-solving strategies, reinforcing students' ability to persist through difficulty. By combining the structured insights of digital games with supportive classroom practices, teachers can create a dynamic environment that not only promotes creativity but also prepares students for complex, real-world problem-solving tasks.

5.3 Potential Improvements to Shadowspect's Game Design

Beyond the current gameplay structure, certain modifications to Shadowspect's design could further support creative engagement. For instance, implementing reward mechanisms for producing multiple distinct solutions could encourage divergent thinking and motivate students to explore alternative strategies. Similarly, incorporating collaborative features, such as allowing students to share and reflect on each other's designs, may stimulate idea generation and peer learning. However, such changes could also introduce challenges, including managing the balance between independent and socially influenced creativity or preventing superficial exploration motivated solely by external rewards. Future research could explore the impact of specific game design interventions on fostering deeper creative problem-solving.

5.4 Limitations

This section outlines the limitations encountered in applying creativity to educational settings and games, particularly in the context of our RGs.

First, discontinuities in curriculum design and organization can limit the applicability of creative aspects identified in digital games to educational settings. These issues within educational institutions can hinder the integration of innovative approaches to teaching geometry and spatial reasoning through puzzle games.

Second, the concept of creativity, while often treated as universal, varies significantly across cultures (Craft, 2003) and is further influenced by social class–based assumptions such as resilience, self-reliance, persistence, and control over one's environment. These assumptions can influence the perception and measurement of creativity, potentially biasing our findings and complicating creativity's integration into educational settings. The lack of a clear definition and guidelines for fostering creativity in curricula, coupled with the dominance of traditional teaching methods and a focus on conventional testing, further limits the scope of applying creative solutions in real-world education (Cachia et al., 2010).

Moreover, the proposed model was developed and evaluated within the context of the Shadowspect puzzle game, providing a solid foundation for identifying creative solutions in serious digital puzzle games. While the methodology demonstrates potential for generalization, each game presents unique mechanics and educational objectives, necessitating further exploration and adaptation to ensure effective application. Additionally, the manual coding process, despite high inter-rater reliability, inherently involves subjective judgments that may influence the interpretation of creativity. The conclusions drawn are consistent with the data and results but rely on the specific context of Shadowspect, warranting additional validation in other environments to confirm their broader applicability. These factors highlight the importance of contextual adjustments and ongoing refinement to maximize the model's applicability and reliability across diverse settings.

Additionally, although we evaluated model performance using precision, recall, and F1 score to account for class imbalance, we did not apply specific mitigation strategies such as oversampling or cost-sensitive learning. Future research can explore techniques like Synthetic Minority Oversampling Technique (SMOTE) or weighted classifiers to improve sensitivity to the minority class (creative solutions).

Furthermore, as more extensive datasets become available, future work could explore the application of deep learning techniques, such as recurrent neural networks (RNNs) or transformer-based models, to capture complex sequential patterns in students' interaction data. While traditional ML models were appropriate given the current dataset size, deep learning approaches may uncover deeper behavioural representations of creativity once larger datasets allow effective training. In addition, the proposed methodology could be extended beyond creativity assessment to evaluate other complex competencies, such as critical thinking, collaboration, or self-regulation, by adapting the coding schemes and predictive models to new learning contexts and tasks.

Finally, a critical limitation involves the ethical dimensions of fostering creativity. While creativity can lead to constructive outcomes, it also has the potential for destructive uses. The ethical dilemma of promoting creativity without inadvertently encouraging harmful uses poses a significant challenge.

6. Conclusions

Assessing creativity remains a complex yet vital challenge in assessment and game-based contexts in education, demanding approaches that are both rigorous and adaptable. In this study, we explored related work on creativity measurements in digital games and, based on these insights, proposed a general model for identifying creative solutions in serious digital puzzle games. This model was subsequently adapted to the geometry-focused puzzle game Shadowspect, where we manually coded solutions across several creativity aspects, such as fluency, flexibility, and surprise. Leveraging the manually labelled dataset, we trained ML models to classify creative solutions, achieving promising results that demonstrate the model's potential to capture nuanced creative behaviours. These findings underscore the viability of combining manual and computational methods to enhance creativity assessment, paving the way for deeper integration of creativity-focused metrics in educational game design.

Building on these findings, in future research we will focus on refining the proposed creativity assessment model (e.g., including additional dimensions of creativity) to address its limitations and broaden its applicability. Adapting the model to diverse educational games and contexts could help establish its generalizability and uncover unique insights across different domains. Furthermore, exploring how guidance and feedback mechanisms influence the relationship between creative exploration and task completion can help bridge the gap between innovative thinking and practical outcomes. Moreover, investigating the long-term impact of game-based approaches on students' problem-solving skills and academic success will be critical for tailoring educational games to support both engagement and creativity. Finally, conducting a case study in which participants self-evaluate their creativity ability and confidence before and after playing could provide additional insights into creativity assessment.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work has been partially funded 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.

References

Almeida, F., & Simoes, J. (2019). The role of serious games, gamification and industry 4.0 tools in the education 4.0 paradigm. *Contemporary Educational Technology*, *10*(2), 120–136. https://doi.org/10.30935/cet.554469

Blanco-Herrera, J. A., Gentile, D. A., & Rokkum, J. N. (2019). Video games can increase creativity, but with caveats. *Creativity Research Journal*, *31*(2), 119–131. https://doi.org/10.1080/10400419.2019.1594524

Boyles, T. (2012). 21st century knowledge, skills, and abilities and entrepreneurial competencies: A model for undergraduate entrepreneurship education. *Journal of Entrepreneurship Education*, *15*(1), 41–56.

Cachia, R., Ferrari, A., Ala-Mutka, K., & Punie, Y. (2010). *Creative learning and innovative teaching: Final report on the study on creativity and innovation in education in the EU member states* (tech. rep.). European Union. https://doi.org/10.2791/52913

Carroll, M., Ruble, M., Dranias, M., Rebensky, S., Chaparro, M., Chiang, J., & Winslow, B. (2020). Automatic detection of learner engagement using machine learning and wearable sensors. *Journal of Behavioral and Brain Science*, *10*(3), 165–178. https://doi.org/10.4236/jbbs.2020.103010

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal*, *17*(1), 37–50. https://doi.org/10.1207/s15326934crj1701_4

Collard, P., & Looney, J. (2014). Nurturing creativity in education. *European Journal of Education*, *49*(3), 348–364. https://doi.org/10.1111/ejed.12090

Craft, A. (2003). The limits to creativity in education: Dilemmas for the educator. *British Journal of Educational Studies*, *51*(2), 113–127. https://doi.org/10.1111/1467-8527.t01-1-00229

Cropley, A. (1997). Creativity: A bundle of paradoxes. *Gifted and Talented International*, *12*(1), 8–14. https://doi.org/10.1080/15332276.1997.11672859

Cropley, A. J. (2000). Defining and measuring creativity: Are creativity tests worth using? *Roeper Review*, *23*(2), 72–79. https://doi.org/10.1080/02783190009554069

Csikszentmihalyi, M., & Wolfe, R. (2014). New conceptions and research approaches to creativity: Implications of a systems perspective for creativity in education. In *The systems model of creativity: The collected works of Mihaly Csikszentmihalyi* (pp. 161–184). Springer. https://doi.org/10.1007/978-94-017-9085-7_10

De-Marchis, G. P., & Shchebetenko, S. (2022). Creativity studies within the European Union: A bibliometric analysis. *Creativity Research Journal*, *36*(1), 116–137. https://doi.org/10.1080/10400419.2022.2067961

Ellison, M., & Drew, C. (2019). Using digital sandbox gaming to improve creativity within boys' writing. *Journal of Research in Childhood Education*, *34*(2), 277–287. https://doi.org/10.1080/02568543.2019.1675823

Erbas, A. K., & Bas, S. (2015). The contribution of personality traits, motivation, academic risk-taking and metacognition to the creative ability in mathematics. *Creativity Research Journal*, *27*(4), 299–307. https://doi.org/10.1080/10400419.2015.1087235

Glaveanu, V. P., Hanchett Hanson, M., Baer, J., Barbot, B., Clapp, E. P., Corazza, G. E., Hennessey, B., Kaufman, J. C., Lebuda, I., Lubart, T., Montuori, A., Ness, I. J., Plucker, J., Reiter-Palmon, R., Sierra, Z., Simonton, D. K., Neves-Pereira, M. S., & Sternberg, R. J. (2019). Advancing creativity theory and research: A socio-cultural manifesto. *The Journal of Creative Behavior*, *54*(3), 741–745. https://doi.org/10.1002/jocb.395

Gomez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2023). A systematic literature review of game-based assessment studies: Trends and challenges. *IEEE Transactions on Learning Technologies*, *16*(4), 500–515. https://doi.org/10.1109/TLT.2022.3226661

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. https://doi.org/10.1016/j.softx.2024.101763

Haylock, D. (1997). Recognising mathematical creativity in schoolchildren. *Zentralblatt für Didaktik der Mathematik*, *29*(3), 68–74. https://doi.org/10.1007/s11858-997-0002-y

Hernández-Torrano, D., & Ibrayeva, L. (2020). Creativity and education: A bibliometric mapping of the research literature (1975–2019). *Thinking Skills and Creativity*, *35*, 100625. https://doi.org/10.1016/j.tsc.2019.100625

Hsiao, H.-S., Chang, C.-S., Lin, C.-Y., & Hu, P.-M. (2014). Development of children's creativity and manual skills within digital game-based learning environment. *Journal of Computer Assisted Learning*, *30*(4), 377–395. https://doi.org/10.1111/jcal.12057

Isaksen, S. G., Stead-Dorval, K. B., & Treffinger, D. J. (2023). *Creative problem solving: An introduction*. Routledge. https://doi.org/10.4324/9781003419327

Khalil, R., & Moustafa, A. A. (2022). A neurocomputational model of creative processes. *Neuroscience & Biobehavioral Reviews*, *137*, 104656. https://doi.org/10.1016/j.neubiorev.2022.104656

Leikin, R. (2009). Exploring mathematical creativity using multiple solution tasks. In *Creativity in mathematics and the education of gifted students* (pp. 129–145). Brill. https://doi.org/10.1163/9789087909352_010

Long, H., Kerr, B. A., Emler, T. E., & Birdnow, M. (2022). A critical review of assessments of creativity in education. *Review of Research in Education*, *46*(1), 288–323. https://doi.org/10.3102/0091732x221084326

Plucker, J. A. (2023). The patient is thriving! Current issues, recent advances, and future directions in creativity assessment. *Creativity Research Journal*, *35*(3), 291–303. https://doi.org/10.1080/10400419.2022.2110415

Rafner, J., Biskjær, M. M., Zana, B., Langsford, S., Bergenholtz, C., Rahimi, S., Carugati, A., Noy, L., & Sherson, J. (2022). Digital games for creativity assessment: Strengths, weaknesses and opportunities. *Creativity Research Journal*, *34*(1), 28–54. https://doi.org/10.1080/10400419.2021.1971447

Rahimi, S., Walker, J. T., Lin-Lipsmeyer, L., & Shin, J. (2024). Toward defining and assessing creativity in sandbox games. *Creativity Research Journal*, *36*(2), 194–212. https://doi.org/10.1080/10400419.2022.2156477

Ruipérez-Valiente, J. A., & Kim, Y. J. (2020). Effects of solo vs. collaborative play in a digital learning game on geometry: Results from a K12 experiment. *Computers & Education*, *159*, 104008. https://doi.org/10.1016/j.compedu.2020.104008

Runco, M. A. (2008). Creativity and education. *New Horizons in Education*, *56*(1), 107–115.

Shaheen, R. (2010). Creativity and education. *Creative Education*, *1*(3), 166–169. https://doi.org/10.4236/ce.2010.13026

Silver, E. A. (1997). Fostering creativity through instruction rich in mathematical problem solving and problem posing. *Zentralblatt für Didaktik der Mathematik*, *29*(3), 75–80. https://doi.org/10.1007/s11858-997-0003-x

Snyder, H. T., Hammond, J. A., Grohman, M. G., & Katz-Buonincontro, J. (2019). Creativity measurement in undergraduate students from 1984–2013: A systematic review. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 133–143. https://doi.org/10.1037/aca0000228

Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning*, *34*(4), 366–377. https://doi.org/10.1111/jcal.12263

Sternberg, R. J. (2019). Enhancing people's creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (pp. 88–104). Cambridge University Press. https://doi.org/10.1017/9781316979839.007

Strukova, S., Marco, R. G., Mármol, F. G., & Ruipérez-Valiente, J. A. (2023). Identifying professional photographers through image quality and aesthetics in Flickr. *Expert Systems*, *41*(4). https://doi.org/10.1111/exsy.13526

Strukova, S., Ruipérez-Valiente, J. A., & Mármol, F. G. (2023). Adapting knowledge inference algorithms to measure geometry competencies through a puzzle game. *ACM Transactions on Knowledge Discovery from Data*, *18*(1). https://doi.org/10.1145/3614436

Torrance, E. P. (1966). Torrance tests of creative thinking. *Educational and Psychological Measurement*.

van Broekhoven, K. (2023). The evaluation and selection of creative ideas in educational settings: Current knowledge and future directions. *Creativity Research Journal*, *37*(2), 307–313. https://doi.org/10.1080/10400419.2023.2253403

Walia, C. (2019). A dynamic definition of creativity. *Creativity Research Journal*, *31*(3), 237–247. https://doi.org/10.1080/10400419.2019.1641787

Wingström, R., Hautala, J., & Lundman, R. (2022). Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists. *Creativity Research Journal*, *36*(2), 177–193. https://doi.org/10.1080/10400419.2022.2107850