Intelligenza Artificiale 18 (2024) 103–119
DOI 10.3233/IA-240030
IOS Press

Automating board-game based learning. A comprehensive study to assess reliability and accuracy of AI in game evaluation

Andrea Tinterri[a,*], Federica Pelizzari[b], Marilena di Padova[c], Francesco Palladino[d], Giordano Vignoli[e] and Anna Dipace[f,1]

[a]*Department of Human Sciences, IUL Telematic University, Florence, Italy*
[b]*Department of Pedagogy, Catholic University of the Sacred Heart, Milan, Italy*
[c]*Department of Humanities, Letters, Cultural Heritage and Educational Studies, Foggia, Italy*
[d]*University of Modena and Reggio Emilia, Modena, Italy*
[e]*IESS, European Institute for Superior Studies, Reggio Emilia, Italy*
[f]*Faculty of Human Sciences, Education, and Sport, Pegaso Telematic University, Naples, Italy*

AUTHOR COPY

Abstract. Game-Based Learning (GBL) and its subset, Board Game-Based Learning (bGBL), are dynamic pedagogical approaches leveraging the immersive power of games to enrich the learning experience. bGBL is distinguished by its tactile and social dimensions, fostering interactive exploration, collaboration, and strategic thinking; however, its adoption is limited due to lack of preparation by teachers and educators and of pedagogical and instructional frameworks in scientific literature. Artificial intelligence (AI) tools have the potential to automate or assist instructional design, but carry significant open questions, including bias, lack of context sensitivity, privacy issues, and limited evidence. This study investigates ChatGPT as a tool for selecting board games for educational purposes, testing its reliability, accuracy, and context-sensitivity through comparison with human experts evaluation. Results show high internal consistency, whereas correlation analyses reveal moderate to high agreement with expert ratings. Contextual factors are shown to influence rankings, emphasizing the need to better understand both bGBL expert decision-making processes and AI limitations. This research provides a novel approach to bGBL, provides empirical evidence of the benefits of integrating AI into instructional design, and highlights current challenges and limitations in both AI and bGBL theory, paving the way for more effective and personalized educational experiences.

Keywords: Board game-based learning, artificial intelligence in education, pedagogical frameworks, educational game design

1. Introduction

1.1. Context of the study

This study explores the implementation of generative artificial intelligence (AI) to facilitate Game-Based Learning (GBL) and its subset, Board Game-Based Learning (bGBL), both of which harness the power of games to enhance the educational experience. The landscape of educational practice is

[1] Conceptualization, A.T. and F.Pe.; introduction and theoretical framework, A.T..; methodology and context of application, F.Pe.; data acquisition, M. di P.; data analysis, A.T., game evaluation G.V., F.Pa. and F.Pe; writing—original draft preparation, A.T. and F.Pe.; writing—review and editing, A.T., M. di P., F.Pa., F.Pe., G.V., and A.D.; supervision, A.D. All authors have read and agreed to the published version of the manuscript.

[*] Corresponding author: Andrea Tinterri, E-mail: atinterri@iuline.it.

ISSN 1724-8035/$35.00 © 2024 – IOS Press. All rights reserved.

evolving towards learner-centered methodologies to develop relevant competencies and prepare today's students, and tomorrow citizens, to a society driven by rapid and unpredictable transformations. As these transformative approaches become more prevalent, they bring with them complex challenges: While the literature on GBL and bGBL presents a compelling case for their effectiveness [1–8], translating these findings into successful implementation in real-world educational settings has proven difficult [9–13]. Educators often struggle with the task of designing and delivering game-based lessons that align with the evidence-based principles outlined in the literature [10, 14–16]. The divide between theory and practice requires careful planning, teacher training, and ongoing support to bridge the gap effectively [17–19]. Within this context, we consider the profound impact of AI in conjunction with GBL. AI has the potential to revolutionize education by enhancing the quality of teaching and learning practices, promoting personalized learning, automating routine tasks and allowing smart management of classes and resources [20–30]. In particular, general large language models (LLM) such as GPT-4 can act as a "swiss knife" to assist a variety of educational processes. However, along with ethical and transparency issues, one significant question that educators and researchers face is the reliability and the quality of responses generated by AI systems in the context of GBL and bGBL. The issue of prompting and the potential for AI models to produce erroneous or incomplete information are valid considerations that can make or break the success of AI-enhanced education [29–31]. Thus, the effectiveness of AI-supported GBL hinges on the ability to provide accurate, relevant, and meaningful guidance to teachers and learners.

AUTHOR COPY

1.2. Game-Based Learning (GBL) and board Game-Based Learning (bGBL): a pedagogical and ludic perspective

Game-Based Learning (GBL) and Board Game-Based Learning (bGBL) are innovative and dynamic educational methodologies that use games as effective tools for fostering learning and knowledge acquisition [32–35]. In GBL digital, video, or tabletop games become a pedagogical medium [36, 37] that can be leveraged to sustain a variety of educational scenarios [38–40]. Games can captivate learners by transforming complex concepts and academic content into enjoyable challenges and quests [41]. As learners navigate the game system, they absorb information, strategize employing critical thinking and problem-solving skills, often without even realizing they are actively engaged in complex learning tasks [42]. This allows for differentiated instruction, accommodating diverse learning styles and abilities, and promoting inclusivity in education [43]. bGBL is a subset of GBL, focusing specifically on tabletop board games [44] as the educational catalyst. Board games offer a tactile and social dimension to the learning experience [45]. Depending on the game, gameplay is driven by interaction between player(s) and the physical components of the game, such as dice, cards, board, or tokens. bGBL promotes face-to-face interaction, collaboration, and healthy competition, fostering interpersonal skills and teamwork among learners [46–48]. Moreover, it often encourages strategic thinking, decision-making, and the application of knowledge in a real-world context [6, 49, 50]. bGBL has specific features that distinguish it from GBL based upon digital (video)games (sometimes referred to as vGBL): First, it does not require any digital device or software, internet access, subscriptions, and/or digital competences to be implemented [51]. Second, due to the physical nature of its components, it is easier to modify or personalize games or components to better align with instructional or educational goals or to improve accessibility [52–54] or as a design challenge for students [38, 55, 56]. Third, board games are transparent systems in that the relationship between the parts (rules and components) is explicit and visible to the player, rather than hidden "under the hood" of the calculator [57], favoring exploration, reflection, and tinkering [58]. Fourth, game rules are implemented by players and not the software; this makes board games more prone to rule misinterpretation, which might hamper the game experience by generating discussion or frustration among players. Fifth, board games have limitations in terms of the number of participants and logistics, whereas vGBL can be easily scaled to reach a wider audience through online platforms. Sixth, being primarily based on face-to-face interaction, bGBL is known to elicit a sense of "togetherness" and "social presence" [46, 48] which can enhance emotional and social engagement towards the activity, contributing to social construction of knowledge [57]. Despite the specific properties and educational potential of board games, the scientific literature on game-based learning has mostly focused on digital games, to the point of sometimes identifying GBL with vGBL. This is problematic from a pedagogic point of view since most GBL instructional models are tailored on vGBL and cannot easily be transferred to bGBL due to its specificities [59]. There are a few notable exceptions, such as Weitze's Smiley Model [60], Sousa's MBGTOTEACH framework [61] and Andreoletti and Tinterri's GDBL ID model [62], that are constructed with the specific needs of bGBL in mind. However, a few critical steps in bGBL design remain loosely investigated. Those include defining the specific goals that can be attained through gameplay and aligning them with the learning goals of the activity [63, 64]; personalizing the game of choice to achieve better constructive alignment with the activity goals [52, 58, 65–67]; evaluating the efficacy of play within bGBL [68]; and evaluating and choosing a board game for the learning activity. Researchers themselves often preselect or construct the games used in experiments [58], [69–71]. Indeed, the choice of the game is both a crucial and complex step in GBL requiring curricular knowledge, practical pedagogical or "school" knowledge, knowledge of board games, and GBL competences on the teachers' side [19]. Thus, clear and comprehensive evaluation criteria are needed to ensure that bGBL serves as a productive and intentional methodology rather than a mere diversion [51, 53, 72–74].

1.3. *Using game choice to explore the synergy of board game-based learning and artificial intelligence*

bGBL and artificial intelligence (AI) is a dynamic convergence that is reshaping the educational landscape in profound ways [75–77]. The experiential nature of games invites active participation, fueling intrinsic motivation, which is a cornerstone of effective knowledge acquisition [78]; AI, on the other side, brings to the table sophistication and personalization opportunities [79]. The use of Artificial Intelligence in the educational context (AIEd) has represented a growing interdisciplinary field since the 1970s to improve course design and expected student outcomes [80]. The sudden diffusion and large-scale popularity of new tools such as ChatGPT, Google Bard, Microsoft Copilot, and others, represented a turning point in educational practices: their user-friendly conversational interfaces significantly lowered the entry barrier to using AI, allowing non-experts, including researchers and educators, to assist a growing range of everyday tasks [27]. Their main features include large-scale language models, in-context learning, and reinforcement learning from human feedback [81]. Even though empirical evidence is still limited, most researchers agree that generative AI tools, when integrated into a pedagogical framework, could improve the effectiveness of the teaching and learning process: they enhance the teachers' ability to discern each learner's unique strengths, weaknesses, and pace of learning, tailoring educational content to suit individual needs [22, 28, 31, 82–84]. Researchers suggest that AI can be used to assist or automate instructional design [85], assuming a variety of roles according to different steps. We argue that the same logic can be applied with bGBL design and we describe potential applications of AI tools in bBGL design according to the GDBL ID model, structured upon the popular Analysis, Design, Development, Implementation, Evaluation (ADDIE) [91] framework. We defined for each step of the ADDIE the phases of GDBL and identified potential roles of AI tools for each phase (Table 1).

AI tools can assist the designer in instructional steps that are common to all teaching and learning methods as well as those that are specific to bGBL, supporting decision-making and helping to circumvent roadblocks by providing context-sensitive analysis and suggestions. However, integrating AI in bGBL design requires teachers to know and understand how AI tools function, how to use such tools in a critical and responsible manner and understand ethical implications of using AI, a set of competencies referred to as AI literacy [88, 89]. ChatGPT and similar models have known limitations and drawbacks that must be carefully considered. Those include, but are not limited, to lack of common sense, limited understanding of context, possible use of biased training data, and lack of emotional intelligence [21, 95]. Furthermore, due to their Transformer architecture, which acts by stochastically computing each next word based on the previous [81, 90], tools like ChatGPT are "black boxes" in that there is no way to access the information they were pre-trained on and, at the same time, it is not possible to know what sources are used to formulate answers. This is a key issue that conflicts with current European regulations, which highlight that AI systems must be "developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their right" [90, p.25]. With those crucial issues in mind, from an instructional design perspective it is therefore important to understand whether

AUTHOR COPY

Table 1
Potential applications of AI tools in bGBL design, based on the ADDIE framework [86] and GDBL ID model [62]

ADDIE	Phase of GDBL model	Description	GBL specific?	Potential role of AI tools	Reference
A(nalysis)	Analysis of needs	Determine students' characteristics and needs	No	Analysis of training needs and training plan	[90]
D(esign)	Learning outcomes	Expected learning outcomes in terms of knowledge, skills, and attitudes	No	Assist definition of expected learning outcomes	[31]
	Game goals	Definition of the goals attainable through gameplay	Yes	Assist definition of expected game goals	–
	Acceptable evidence	Quantitative and qualitative data that can be obtained to assess learning outcomes during and after gameplay	No	Concept checking and exam preparation; Drafting assistance.	[25]
	Game shortlist	Selection of a shortlist of potential games based on desired characteristics	Yes	Propose a list of games according to desiderata	–
D(evelopment)	Game choice and personalization	Selection and modification of the game based on alignment with context and design goals	Yes	Analyze and evaluate potential games; propose personalizations	[87]
	Instructional strategies	Define the lesson plan	No	Curate personalized learning paths	[90]
I(mplement)	Implementation	Realize the activity in the classroom	No	Content generation and translation; real-time feedback to students (Chatbots)	[29, 92]
E(valuate)	Evaluation and revision	Evaluate the impact of the activity and revise	No	Providing feedback to students; Automated assessment	[20]

AUTHOR COPY

general LLM-based AI tools can provide sufficient advantages to warrant their implementation in bGBL. To this aim, we defined three key issues that must be addressed. The first is consistency: given their stochastic nature, these tools able to provide consistent output or is it excessively variable to be relied upon? (RQ1). Moreover, the success of AI and bGBL synergy rests on the ability of AI tools to provide not only consistent but also accurate answers to the designer's questions. We are still in an exploratory phase of understanding how accurate artificial general intelligence answers can be; this is especially true for complex, open-ended tasks, such as those are required for the design of bGBL, whereas assessing the accuracy of the AI tools often relies on expert evaluation from humans [92]. Thus, the second question is: are AI tools able to provide sufficiently accurate game suggestions compared to human experts? (RQ2). First indications suggest that the accuracy of the performance depends on the type of challenge [93–95] and the subject [25], as well as from contextual and reinforcement information provided by the user [29, 96–98]. Still, despite the emergence of good practices such as defining a persona, structuring key ideas, and using specific rather than general requests [29, 96, 98, 99], it is not yet clear how contextual and reinforcement input should be organized to maximize output quality and how sensitive the instrument is to user-provided contextual information. This is another key requirement for AI-assisted bGBL: given specific instructions, are AI suggestions sensible to the educational context? (RQ3). Thus, the goal of this exploratory study is to evaluate the consistency, accuracy, and context-sensitivity of ChatGPT as an assisting tool for instructional designers in Board Game-Based Learning (bGBL). To this aim, we tested ChatGPT performance on a specific phase of bGBL, the choice of a board game for the educational activity (Table 1); the rationale for this choice is threefold: a) This process is specific to GBL (Table 1) b) it is critical for successful bGBL [61]: board games are characterized by different game mechanisms [100], structures [101], genres, and themes [102] that allow different kinds of cognitive, emotional, social and motivational engagement [46, 62], favor different pedagogical approaches [103], provide different challenges for accessibility and inclusion [53], and opportunities for internal assessment of educational goals [68]. The process of choosing board games, whether for educational enrichment or sheer entertainment, necessitates a meticulous evaluation of a multitude of factors. c) Game choice is often considered by teachers as one of the most difficult steps in GBL [10, 12, 14, 16, 104] as they often lack the knowledge of games and the ability to achieve constructive alignment [65] between the learning goals and the game activity. When making selections, it's essential to weigh elements like the intended age group, the number of players, the intricacy of game mechanics, alignment with predefined learning objectives, and the desired depth of engagement. Thus, automating this process with generative AI could help overcome one of the main roadblocks that hampers successful implementation of bGBL. Therefore, the research questions for this study have been defined as such:

RQ1: The first research question concerns the consistency of the tool. Are ChatGPT game choices for a specific learning activity replicable, given the same instructions?

RQ2: The second research question concerns the accuracy of the tool. Are ChatGPT game choices comparable with those provided by human experts?

RQ3: The third research question considers the context-sensitivity of the tool. Is ChatGPT able to adapt its game choices according to different didactic backgrounds and/or disciplines?

2. Materials and methods

To answer these questions, the research was conducted in four steps: definition of the instructional framework and contextual information; prompt building and execution; independent expert evaluation; analysis and data comparison (Fig. 1).

2.1. Definition of the instructional framework and contextual information

In educational design, crafting meaningful and impactful learning experiences requires a deliberate and harmonious alignment of various critical components [100]. These components include specific academic disciplines, dimensions of competence, the infusion of engaging and immersive playful scenarios, and the careful selection of pedagogical approaches. This alignment is the keystone upon which the effectiveness of educational interventions is built [101]. We initially defined three "sets" of contextual elements, each instrumental in providing an in-depth understanding of the intricate interplay between these components and how they manifest in diverse learning environments [102].

1. **Background:** This column refers to the educational level or grade, indicating which grade or level the games are intended for, as well as a descriptive profiling of the class, including number and specificities of pupils. Two backgrounds are included in this study: "Primary - III" and "Secondary - I," representing different educational levels and student profiles. The decision to choose both a primary school class and a secondary school class reflects a deliberate effort to capture the diversity of learners' needs and abilities across different educational levels. Primary school students, often at the foundational stage of their learning, require a different approach to engage with educational content compared to secondary school students, who are typically exposed to more advanced and nuanced subject matter.
2. **Disciplines and dimensions of competence:** The column "discipline" specifies the subject or discipline in which the games are being used. In this study, "Italian - Fiction" and "Math - Calculation" were used. The column "Dimensions of Competence" describes the specific skills or competencies that students are expected to develop by using the chosen games within the disciplinary setting. In the primary level example, students are expected to "Recognize various types of text and their characteristics based on given patterns." In the secondary level, students are supposed to "Connect measurement practices to knowledge about numbers and operations."
3. **Educational purpose and pedagogical approach:** This column defines the main educational purpose according to which the game is played in the context of the activity. We employed the GBL framework developed by Thorsten Hangøj [39] to account for the different goals and purposes of GBL from a teachers' perspective. The author refers to the work of

AUTHOR COPY

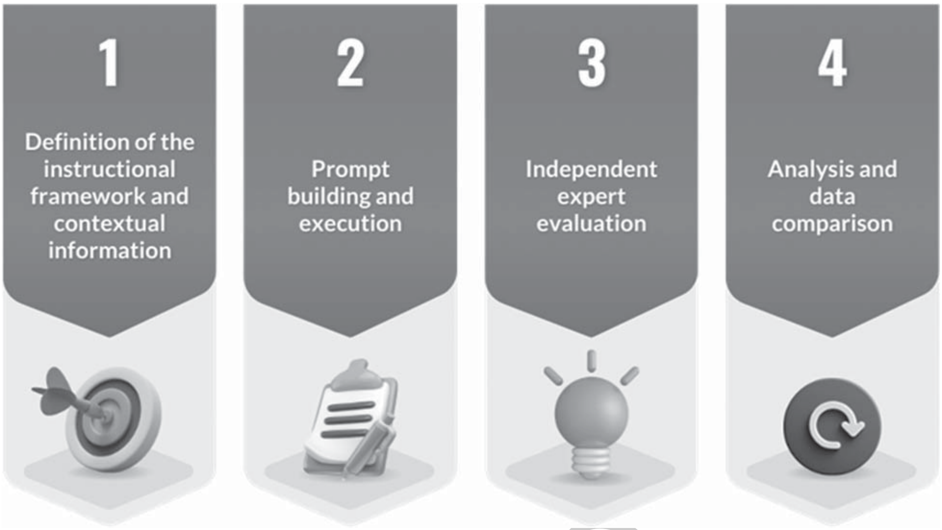
1
Definition of the instructional framework and contextual information
2
Prompt building and execution
3
Independent expert evaluation
4
Analysis and data comparison

Fig. 1. Summary of the experimental phases.

AUTHOR CO

Gert Biesta [59], who defines three fundamental purposes of education, namely qualification, socialization, and subjectification, to identify the possible educational purposes of a GBL scenario. Each of these purposes plays a unique role in shaping the educational process and the development of learners, and evidence suggests that board games can contribute towards such aims:

(a) **Qualification** aims to equip learners with knowledge, skills, and attitudes to perform specific tasks. Board games allow the acquisition of domain-specific knowledge [48], executive functions [4], and high-level skills such as computational thinking [50] or management [71].

(b) **Socialization** aims to integrate individuals into specific social, cultural, and political contexts. Board games allow the development of social and collaborative skills [48, 57].

(c) **Subjectification** fosters independence and individual identity outside established social norms. Board games can foster metacognitive reflection [58], promote speaking skills, and might improve self-mastery [49].

In Hangøj's model, the teacher can adopt different pedagogical approaches that can significantly impact the learning experience in GBL:

1. **Ludic:** The teacher primarily sees games as a fun experience where students find opportunities for self-expression. Emphasizes enjoyment and creative expression and its main purpose is to make learning enjoyable.
2. **Drill and skill:** The teacher views the game experience as an opportunity for the production and reproduction of knowledge through repetition. It reinforces existing knowledge and skills and its aim is skill-building.
3. **Pragmatic:** The teacher approaches the game to simulate aspects of real problems or situations, developing situated knowledge. It helps connect game environments to real-world applications and encourages practical application.

For this study, the "Drill and Skill" and "Pragmatic" approaches are linked to the "Socialization" purpose, whereas the "Drill and Skill" and "Explorative" approaches are tied to the "Qualification" purpose. The eight Learning Units (LU) obtained by combining the three "sets" that constitute the pedagogical framework used in this case study are shown in Table 2.

The choice of competence dimensions, playful scenarios, and pedagogical approaches in educational settings is closely related and should be carefully aligned to create an effective and engaging learning experience. Here's how they are interconnected:

1. **Alignment:** The competence dimensions are the learning objectives. The educational purpose and pedagogical approaches provide a context in which these objectives can be practiced and applied. For instance, in the "PRIMARY - III"

Table 2
Outline of the combinatorial approach for the definition of the Learning Units (LU) used in this case study. Each LU is a unique combination of the three "sets" that constitute the pedagogical framework developed for the study

SET 1	SET 2	SET 3	Background	Disciplines	Dimensions of competence
Primary - III	ITALIAN - FICTION	Recognize various types of text and their characteristics based on given patterns.	Qualification	Drill and Skill/Explorative	LU1
Primary - III	ITALIAN - FICTION	Recognize various types of text and their characteristics based on given patterns.	Socialization	Drill and Skill/Pragmatic	LU2
Secondary - I	ITALIAN-FICTION	Recognize various types of text and their characteristics based on given patterns	Qualification	Drill and Skill/Explorative	LU3
Secondary - I	ITALIAN-FICTION	Recognize various types of text and their characteristics based on given patterns	Socialization	Drill and Skill/Pragmatic	LU4
Secondary - I	MATH - CALCULATION	Connect measurement practices to knowledge about numbers and operations.	Qualification	Drill and Skill/Explorative	LU5
Secondary - I	MATH - CALCULATION	Connect measurement practices to knowledge about numbers and operations.	Socialization	Drill and Skill/Pragmatic	LU6
Primary - III	MATH - CALCULATION	Connect measurement practices to knowledge about numbers and operations.	Qualification	Drill and Skill/Explorative	LU7
Primary - III	MATH - CALCULATION	Connect measurement practices to knowledge about numbers and operations.	Socialization	Drill and Skill/Pragmatic	LU8

level, students are learning to recognize various types of text. The educational purpose of "Socialization" can be used to encourage students to practice this skill by engaging in social interactions that involve different types of text.

2. **Pedagogical strategies:** The pedagogical approaches describe how the educational content is delivered. The approaches should be chosen to support the achievement of competence dimensions according to the main educational purpose. In the "Secondary - I" level, the "Explorative" pedagogical approach can be used to encourage students to explore mathematical concepts with the purpose of "Qualification," where they connect measurement practices to numbers and operations.
3. **Engagement and relevance:** The use of board games adds an element of fun and relevance to the learning process. When competence dimensions are taught within a context that students find enjoyable and meaningful, it can lead to better engagement and retention of knowledge and skills.
4. **Assessment:** The choice of pedagogical approaches can also impact how students' progress is assessed. The assessment methods should align with the chosen competence dimensions and the playful scenario. This ensures that students are evaluated based on their ability to apply their knowledge and skills in a context relevant to the playful scenario.

In the context of the research at hand, a careful and judicious approach was taken to curate a selection of five board games based on a rigorous set of criteria. The selection process focused on games that not only enjoyed widespread recognition and acclaim but were also well-understood and endorsed by experts in the field. The researchers undertook a detailed examination of the main characteristics of chosen games [62,103], summarized below according to the main characteristics of board games: Title, theme, structure, genre, and main mechanisms of the game (Table 3).

To evaluate the effectiveness of games as educational tools, it becomes imperative to establish a set of evaluation criteria. In our illustrative case, we have defined five such criteria:

- Accessibility provided internally by the game: Measures how easy it is for students to access the resources and rules of the game. This metric not only measures the ease of entry into the educational gaming experience but also determines how readily learners can engage with its components.
- Setting (time needed to prepare and play the game): Measures how long it takes to start and complete a game. This factor is vital in understanding whether the game can be seamlessly integrated into the curriculum without consuming excessive time, ensuring that it aligns with educational objectives.

AUTHOR COPY

Table 3
Characteristics of the board games examined. This table offered insights into the essential characteristics of each game, including thematic elements, activated game mechanics, genre categorization, and internal structural intricacies, as defined in [62]

Game	Theme	Structure	Genre	Mechanism
Carcassonne	Medieval	Multiplayer	Family, Eurogame	Tile placement, area control
Codenames	Espionage	Two-team game	Party game, Family	Word association, deduction, team-based play
Concept	Abstract; Communication	Multiplayer	Party game, Family	Deduction, pictorial communication
Dixit	Abstract, Imagination	Multiplayer	Party game, Family	Storytelling, voting
Pandemic	Global Epidemic	Cooperative	Strategy, Family	Hand management, set collection, action points

AUTHOR COPY

– Content (use of thematic knowledge to address the challenges of the game or mere accessory): Examines whether the thematic content of the game contributes to learning or is only a superficial aspect.
– Learning curve Assesses how easy it is for learners to learn to play the game and develop skills over time. It is pivotal in determining whether the game caters to a broad spectrum of learners, from novices to those seeking more advanced challenges, or if it alienates a significant portion of its intended audience.
– Opportunities to assess standards of success: Determines if the game provides opportunities for students to assess their success against the learning objectives and competence dimensions. This feature assesses the game's capability to foster self-assessment and alignment with educational goals.

These criteria serve as a shared yardstick, both for experts in the field and for ChatGPT, when ranking the usefulness of games within various educational contexts. The use of a scale from 1 to 5 empowers us to quantitatively assess the efficacy of these games in the educational domain. Starting from this framework and these evaluation criteria, the experimentation on ChatGPT and with experts was then activated.

2.2. *Prompt building and execution*

The researchers first set up a prompt to obtain game rankings from ChatGPT as follows: the first prompt set up the persona [98], provided specific information concerning the task [96] and, using a numbered list, indicated the content that would be provided in successive prompts, including the definition of playful scenarios, pedagogical approaches, the board games to be ranked, the criteria for board game evaluation, and the two educational contexts. Furthermore, we provided in successive prompts the characteristics of each LU. A summary of the prompts used to obtain answers from ChatGPT are presented in Fig. 2.

2.3. *Independent expert evaluation*

The collaborative evaluation of bGBL strategies featured the integration of ChatGPT with a panel of three expert evaluators [104], each of whom possessed a wealth of domain-specific knowledge and extensive experience in the realm of instructional design and bGBL. These experts were selected to bring their unique insights and specialized understanding to the assessment process, ensuring a comprehensive and informed evaluation. The experts were provided with the same information (educational context, description of the framework and clarification of the main terms) that were inserted in ChatGPT prompts. The evaluation process entailed the use of a blind Excel file, which served as a standardized platform for assessing the bGBL strategies. Within this Excel file, experts had to rank, for each LU, the games from least to most suitable for the activity, as well as providing a short motivation for their choices based on assessment criteria. This approach allowed for a clear and systematic comparison of the strategies.

2.4. *Analysis and data comparison*

To measure the internal consistency of ChatGPT (RQ1), we pooled the rankings given to LUs for pairwise comparison. We first analyzed the correlation between the results obtained by three single runs on the same account. The first two runs were obtained by using the "regenerate answer" option. The third run was obtained by running a different chat, with the same prompts as the former. To measure external consistency, we ran this process on two different

[USER] You are a curricular development expert system focused on board Game-based learning (bGBL). bGBL is a learning strategy aimed at promoting knowledge through the use of board games in the classroom. Your goal is to rank a set list of boardgames according to how well they suit learning units for which the user will provide a) educational context b) subject matter c) dimensions of competence to be developed d) playful scenarios e) pedagogical approaches f) select games. The user will provide: 1) a definition of playful scenario 2) a definition of pedagogical approaches 3) a set of five commercial board games 4) criteria for evaluating games 5) two educational contexts, one called [Primary III] and the other [Secondary I].
[USER] 1) Definition of Playful Scenarios [1077 characters].
[USER] 2) Definition of Pedagogical approaches [590 characters]
[USER] 3) The set of board games to rank include: Carcassonne, Codenames, Concept, Dixit, Pandemic.
[USER] 4) Criteria for evaluation: Accessibility and inclusion. Setting. Content. Learning Curve. Opportunities to assess success standards.
[USER] 5) [Primary III] [3382 characters]
[USER] [Secondary I] [3583 characters]
[USER] The user will provide eight combinations [only four were retained for this study, nda] of educational context, subject matter, dimensions of competence to be developed, playful scenarios, pedagogical approaches. For each combination, you will provide a rubric assigning numerical scores from 1 to 5 according to the following dimensions [574 characters]. Then, rank boardgames from the most to the least indicated for the learning unit. LU1: a) [Primary III] b) Italian - fiction c) Recognize various types of text and their characteristics based on given patterns. d) Qualification e) Drill and Skill/Explorative
[USER] LU2: a) [Primary III] b) Italian - fiction c) Recognize various types of text and their characteristics based on given patterns. d) Socialization e) Drill and Skill/Pragmatic
[USER] LU3: a) [Secondary I] b) Italian - fiction c) Recognize various types of text and their characteristics based on given patterns. d) Qualification e) Drill and Skill/Explorative
[USER] LU4: a) [Secondary I] b) Italian - fiction c) Recognize various types of text and their characteristics based on given patterns. d) Socialization e) Drill and Skill/Pragmatic
[USER] LU5: a) [Secondary I] b) Math-Calculation c) Connect measurement practices to knowledge about numbers and operations. d) Qualification e) Drill and Skill/Explorative
[USER] LU6: a) [Secondary I] b) Math-Calculation c) Connect measurement practices to knowledge about numbers and operations. d) Socialization e) Drill and Skill/Pragmatic
[USER] LU7: a) [Primary III] b) Math-Calculation c) Connect measurement practices to knowledge about numbers and operations. d) Qualification e) Drill and Skill/Explorative
[USER] LU8: a) [Primary III] b) Math-Calculation c) Connect measurement practices to knowledge about numbers and operations. d) Socialization e) Drill and Skill/Pragmatic

Fig. 2. The table shows the user messages provided to ChatGPT to obtain game rankings for each LU.

accounts. We averaged the results obtained by the three runs on the same account, obtaining an ensemble (a new ranking based on the averaged results); then, we compared the ensembles obtained by the two accounts. To provide a measure of consensus between ChatGPT and between the experts' evaluation (RQ2), we first measured disagreement between individual experts (Table 6). Then, we pooled together experts ranking for each LU and compared it with the Ensemble ranking obtained from ChatGPT. Then, we pooled together experts ranking for each LU and compared it with the Ensemble ranking obtained from ChatGPT. We used Spearman's Rank correlation and Kendall's Tau correlation; both are commonly used rank-based coefficients used to assess monotonic relationship between ordinal data. Spearman's method measures the pairwise disagreements between two rankings [105], whereas Kendall's Tau focuses on the concordance between pairs of observations [106]. The test output is in both cases a value that lies on a scale from minus 1 to 1,where values of 1, minus 1, and 0 signify a perfect positive relationship, a perfect negative relationship, and no overall ordinal relationship at all, respectively [107]. Generally, values between 0.9 and 1 are considered very high, between 0.7 and 0.9 high, and values between 0.4 and 0.7 moderate positive relationships [108]. We also used Kemeny distance, a measure based on the assumption that voters have the same probability of comparing correctly two alternatives [109], to measure disagreement as the minimum number of pairwise swaps needed to transform one ranking into the other. This approach enables cross-validation of results, enhancing the reliability and validity in assessing the consistency between game rankings. To measure the context-sensitivity of ChatGPT and experts (RQ3), we matched the ensemble rankings of both ChatGPT and experts' rankings according to the dimension investigated. We formed pairs of LUs to isolate and investigate specific dimensions (e.g. to investigate the "background" dimension, LU1 was paired with LU3, LU2 was paired with LU4, and so on; to investigate the "discipline" dimension, LU1 was paired with LU7, LU2 with LU8, and so on; etc.) and used the degree of discordance among ratings as an indicator of context-sensitivity. Statistical analysis was performed using Google Colab; ChatGPT was used to assist the generation of Python code, which was tested and verified by an independent expert.

3. Results

We first analyzed the internal consistency of the answers provided by ChatGPT. To this aim, for each LU we calculated the ensemble score by averaging the repeat rankings obtained within the same account, repeating the analysis for two different accounts (Ensembles 1 and 2). For each LU, we show the average ranking (with standard deviation in parentheticals) and the Kemeny distance indicating total disagreement within each account (Table 4).

AUTHOR COPY

Table 4
The table shows the ensemble rankings for two ChatGPT accounts. Ensemble 1 is the average of runs 1-3, Ensemble 2 is the average of runs 4-6. Kemeny distance shows disagreement between pairs of runs (Run 1-Run 2; Run 1-Run 3; Run 2-Run 3 for Ensemble 1; Run 4-Run 5; Run 4-Run 6; Run 5-Run 6 for Ensemble 2) which are indicated between parentheticals

Context	GPT Ensemble	Carcassonne	Codenames	Concept	Dixit	Pandemic	Kemeny distance (d)
LU1	1	2 (±0)	4 (±1)	3.7 (±0.6)	4 (±0.6)	1 (±0)	(2+4+4) = 10
	2	2 (±0)	3.3 (±0.6)	4 (±1)	4.7 (±0.6)	1 (±0)	(4+2+2) = 8
LU2	1	2 (±0)	4.3 (±0.6)	3 (±0)	4.7 (±0.6)	1 (±0)	(2+2+0) = 4
	2	2 (±0)	3.7 (±0.6)	3.3 (±0.6)	5 (±0)	1 (±0)	(0+2+2) = 4
LU3	1	2 (±0)	5 (±0)	3.3 (±0.6)	3.7 (±0.6)	1 (±0)	(0+2+2) = 4
	2	2 (±0)	3.7 (±0.6)	3.3 (±0.6)	5 (±0)	1 (±0)	(0+2+2) = 4
LU4	1	2 (±0)	5 (±0)	3.7 (±0.6)	3.3 (±0.6)	1 (±0)	(2+0+2) = 4
	2	2 (±0)	4 (±0)	3 (±0)	5 (±0)	1 (±0)	0
LU5	1	5 (±0)	2.7 (±0.6)	3 (±1)	1.3 (±0.6)	3 (±1.7)	(2+6+6) = 14
	2	5 (±0)	3 (±0)	2.3 (±0.6)	1.3 (±0.68)	3 (±1.7)	(6+6+0) = 12
LU6	1	4.3 (±0.58)	2.3 (±0.6)	2.7 (±0.6)	1 (±0)	4.7 (±0.6)	(0+4+4) = 8
	2	5 (±0)	3 (±0)	2 (±0)	1 (±0)	4 (±0)	0
LU7	1	5 (±0)	2.3 (±0.6)	3.3 (±0.6)	1.3 (±0.6)	3 (±1.7)	(0+6+6) = 12
	2	5 (±0)	3 (±0)	2 (±0)	1 (±0)	4 (±0)	0
LU8	1	5 (±0)	2.3 (±0.6)	3.3 (±0.6)	1.7 (±0.6)	3.7 (±1.7)	(0+6+6) = 12
	2	5 (±0)	3 (±1)	2 (±0)	1 (±0)	4 (±0)	0

AUTHOR COPY

For Ensemble 1, we observed strong agreement on game rankings when using the "Regenerate answer" (run 1-2, tau = 0.91, $p < 0.001$; rho = 0.95, $p < 0.001$, d = 8) and moderate agreement when comparing output from two different chats (run 1-3, tau = 0.57, $p > 0.001$; rho = 0.642, $p < 0.001$; d = 31; run 2-3, tau = 0.57, $p = 0.001$; rho = 0.64, $p < 0.001$; d = 31). In Ensemble 2, we observed strong agreement both within the same chat (run 4-5, tau = 0.81, $p < 0.001$; rho = 0.86, $p < 0.001$; d = 12) and different chats (run 4-6, tau = 0.83, $p < 0.001$, rho = 0.88, $p < 0.001$, d = 13; run 5-6, tau = 0.88, $p < 0.001$; rho = 0.93, $p < 0.001$; d = 9). Thus, we observed significant positive agreement in repeated runs from the same account, going from moderate to very strong, indicating good internal consistency of ChatGPT answers. For external consistency, we compared the agreement between Ensemble 1 and Ensemble 2. We found significant, moderate agreement (tau = 0.67, $p < 0.001$; rho = 0.77, $p < 0.001$; d = 26) between the two ensembles, suggesting that, when provided the same prompt, repeated ChatGPT answers are sufficiently consistent across different accounts but subject to variability. The second research question aimed to understand whether ChatGPT rankings were accurate compared to human experts. To this aim, three external evaluators, all competent with the five games provided, were provided with the same information as ChatGPT and asked to rank the games anonymously and independently according to the different combinations. There was significant, moderate agreement among evaluators' rankings (Evaluator 1-Evaluator 2: tau = 0.48, $p < 0.001$; rho = 0.56, $p < 0.001$; d = 34; Evaluator 2-Evaluator 3: tau = 0.37, $p = 0.004$, rho = 0.43, $p = 0.005$, d = 39; Evaluator 1-Evaluator 3: tau = 0.68, $p < 0.001$, rho = 0.77 $p < 0.001$, d = 23). However, some context/scenario combinations elicited more disagreement than others: for instance, in Primary III/Italian-Fiction, all three evaluators chose a different game as first choice: Concept, Dixit and Codenames received one first-place ranking each; in general, LUs with Primary III context elicited more disagreement from evaluators than Secondary I. The mean rankings obtained for each LU (standard deviations in parentheticals), along with the Kemeny distance indicating total disagreement between evaluators, are shown in Table 5.

From the average ranking of the evaluators, we obtained a pooled ranking for each LU and compared it with the Ensemble ranking obtained from ChatGPT. A comparison of the pooled rankings obtained by AI and evaluators is shown, together with Kemeny distance measuring total disagreement (Table 6). Overall, ChatGPT rankings have a moderate positive correlation with experts ranking (tau = 0.477, $p < 0.001$; rho = 0.59, $p < 0.001$), indicating that ensembled ChatGPT rankings are reasonably accurate with respect to expert evaluation, although accuracy varies according to the specific LU.

The third research question concerns the sensitivity of ChatGPT to contextual information. For

Table 5
The table compares the mean rankings from bGBL experts for the eight LUs examined. Data are presented as mean ± standard deviation. Kemeny distance shows total disagreement between individual raters (Evaluator 1-Evaluator 2; Evaluator 1-Evaluator 3; Evaluator 2-Evaluator 3), indicated in parentheticals

Context	Carcassonne	Codenames	Concept	Dixit	Pandemic	Kemeny distance (d)
LU1	1.7 (±0.6)	3.3 (±1.6)	4.3 (±0.6)	4 (±1)	1.7 (±1.2)	(4+6+8) = 18
LU2	1.3 (±0.6)	4.7 (±0.6)	3.7 (±0.6)	2.7 (±0.6)	2.7 (±2.1)	(2+6+8) = 16
LU3	1 (±0)	3 (±0)	5 (±0)	2.7 (±0.6)	3.3 (±1.2)	(4+4+0) = 8
LU4	1 (±0)	4.3 (±1.2)	3.7 (±1.2)	2 (±0)	4 (±0)	(4+0+4) = 8
LU5	5 (±0)	3 (±0)	1.3 (±0.6)	1.7 (±0)	4 (±0)	(2+0+2) = 4
LU6	4.7 (±0.6)	3.3 (±0.6)	1.3 (±0.6)	1.7 (±0)	4.3 (±0.6)	(4+3+3) = 10
LU7	5 (±0)	2 (±1.7)	3 (±0)	2 (±0)	3 (±1.7)	(6+0+6) = 12
LU8	4.3 (±0.6)	3 (±1)	2.3 (±1.2)	1.7 (±0.6)	3.7 (±2.3)	(8+4+8) = 20

Table 6
The table compares the pooled rankings from ChatGPT and the pooled rankings of bGBL experts for the four didactic contexts examined. Kemeny distance shows total disagreement between the pooled rankings from ChatGPT and bGBL experts for each LU

Context		Carcassonne	Codenames	Concept	Dixit	Pandemic	Kemeny distance (d)
LU1	ChatGPT	2	3	4	5	1	3
	Evaluators	2	3	5	4	2	
LU2	ChatGPT	2	4	3	5	1	7
	Evaluators	1	5	4	3	3	
LU3	ChatGPT	2	5	3	5	1	10
	Evaluators	1	3	5	2	4	
LU4	ChatGPT	2	5	3	4	1	6.5
	Evaluators	1	5	3	2	4	
LU5	ChatGPT	5	4	2	1	4	3
	Evaluators	5	3	1	2	4	
LU6	ChatGPT	5	3	2	1	4	2
	Evaluators	5	3	1	2	4	
LU7	ChatGPT	5	2	2	1	4	3
	Evaluators	5	2	4	2	4	
LU8	ChatGPT	5	3	2	1	4	0
	Evaluators	5	3	2	1	4	

the "Background" dimension, ChatGPT paired LUs show a moderate positive correlation (tau = 0.56, $p = 0.002$, rho = 0.64, $p < 0.003$, d = 14), indicating that the software slightly adjusted its ranking according to the background; for comparison, experts show a similar tendency, although with a slightly lower correlation (tau = 0.42, $p = 0.022$, rho = 0.518, $p = 0.02$, d = 20). For the "discipline" dimension, ChatGPT shows a moderate negative correlation (tau = –0.39, $p < 0.03$; rho = –0.54, $p = 0.01$; d = 44) indicating that its ranking decisions were strongly dependent on the discipline, so that games that were ranked high for Italian tended to receive a low rating for Mathematics, and vice versa. Experts ranking also show a similar tendency, although in this case the negative correlation was not statistically significant (tau = –0.32, $p = 0.08$; rho = 0.41, $p = 0.07$; d = 36). Finally, for the "educational purpose/pedagogical approaches" dimension, ChatGPT showed a strong correlation between matched LUs (tau = 0.726, $p < 0.001$; rho = 0.797, $p < 0.001$, d = 10) suggesting that its rankings were only slightly influenced by this dimension. Once again, this tendency is similar to the behaviour of human evaluators, whose ratings for LUs differing for the playful scenario/pedagogical approach show moderate to strong correlation (tau = 0.59, $p < 0.001$; rho = 0.705, $p < 0.001$; d = 14). Taken together, these results suggest that both ChatGPT and human evaluators modify their ranking according to contextual factors, that they seem to weight similarly the influence of different dimensions in their rankings, and that among the dimensions examined in this study "discipline" seem to be by far the most influential for both human evaluators and ChatGPT, although experts seem to factor in slightly more the other dimensions. These results are summarized in Fig. 3.

AUTHOR COPY

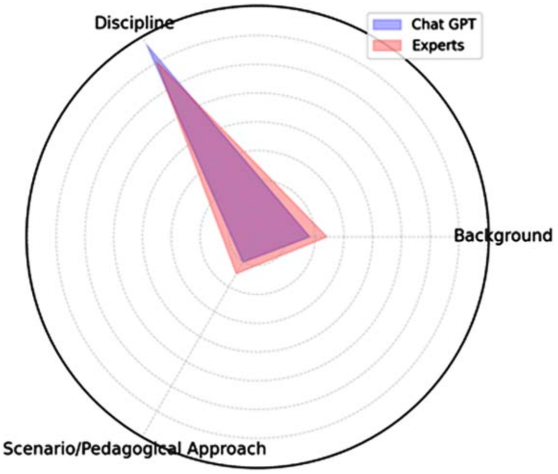

Fig. 3. A visual indication of the influence of contextual elements in board game rankings by ChatGPT and experts. The graph represents Spearman's rank correlations from 1 (origin) to -1 (outer circle). The higher the correlation, the lower the estimated incidence of the dimension on the ranking decision.

4. Discussion

In this study, we analyzed crucial properties of ChatGPT for the task of automating board game-based learning (bGBL): consistency (RQ1), accuracy (RQ2), and context-sensitivity (RQ3). We found that the current version of ChatGPT is not entirely consistent. Given the same prompts, the software tends to provide overall similar answers within the same chat, across different chats and different accounts. Still, there is a certain degree of variability that cannot be explained within the confines of the current study. It is important to acknowledge the limitations and fallibility of AI tools to "address the impact of dogmatic overconfidence in possibly erroneous suggestions" generated by large language models [97]. The potential inconsistency of AI-generated answers is a significant problem for the automation of bGBL design; further research is needed to understand whether the variability observed can be addressed by reducing the complexity of open tasks, for instance by providing more accurate and analytic prompts, in addition to generate ensemble answers by repeating prompts in different conditions (regenerating answers, running separate chats, repeating the prompts on multiple accounts). Due to the design of the study, this result is subject to several limitations: for instance, we limited the analysis to only one general AI tool, Open AI's ChatGPT with GPT-4 model. It is likely that using different models, such as GPT-3.5, Anthropic Claude, or Google Bard could have led to significantly different results. Furthermore, GPT-4 being a premium, paid-for tool, coupled with its "black-box" nature, suggest that alternative AI models more in line with European regulations on the use of AI in education [91] should be explored in the future. Second, our study indicates that, when provided with similar background information, ChatGPT can produce board game recommendations that are comparable with those of bGBL experts. The evaluation of a game for a bGBL activity is a complex, open-ended decision; as would be expected given the difference in experience, background, and individual preferences, even experts' rankings were consistent but not overlapping, as each expert considers and values the given criteria differently. Still, when comparing the pooled answers from both experts and ChatGPT, we observed a significant correlation, suggesting that ChatGPT can be considered accurate enough to assist the choice of games for bGBL activity. To further validate this claim, we investigated how the choices of the chatbot were influenced by the instructional context and examined how it compared with human evaluators. Context was provided through a combinatorial set of didactic backgrounds, disciplinary contents, and educational purposes coupled with the relative pedagogical approaches. By matching the rankings for contexts (LUs) that differed for only one of those factors, we observed that the software can modify its choices according to the variables provided. Although we cannot exclude that the internal variability of ChatGPT accounted for part of the differences observed, the extent to which the rankings varied according to contextual factors was surprisingly similar to that of human experts (Fig. 3). Game evaluations were strongly influenced by the discipline dimension, whereas the class background and the scenario/pedagogical dimensions altered evaluators' ranking less dramatically. However, it is important to highlight the experimental constraints that limit the generalization of these results. First, given the still-evolving literature on prompting techniques, and the exploratory nature of the study, the construction of prompts might have introduced biases in ChatGPT answers. In particular, the background information was provided via two long, descriptive prompts (Fig. 2), whereas information concerning the pedagogical scenarios and approaches was more analytic and information on the discipline was provided very synthetically. Since we currently have little knowledge on how ChatGPT incorporates contextual information produced by the user, we cannot exclude that this influenced

AUTHOR COPY

AUTHOR COPY

the results obtained. A limitation of this study is that we did not compare results produced with different prompting techniques. Second, the need for competent external evaluation to assess the accuracy of AI choices influenced the experimental setup: all experts were experienced in using GBL in school, orienting the definition of background to real school cases, and limiting the choice to five games that all experts were competent with. Codenames, Concept and Dixit are all family-oriented, language-based games, whereas Carcassonne and Pandemic focus on different processes (spatial reasoning and cooperative decision-making, respectively) and share considerably less similarities than the three former games. This might have contributed to the bias towards the discipline dimension observed in game ranking. Furthermore, given that all games share a similar level of complexity (with the partial exception of Pandemic, slightly more complex than the other games), this might have limited the influence of background factors. Further studies will be needed to assess ChatGPT ability to select and evaluate games when given a wider range of educational contexts and less restrictive choices. A third issue is the lack of evidence and guidance in the literature for critical aspects of GBL, and bGBL, instructional design: current limitations include lack of: a) a clear definition of the multifaceted educational goals that can be achieved with games, b) guidance towards the achievement of constructive alignment between learning goals and game goals, c) criteria for selecting and evaluating specific games, as well as the related question of linking specific board game mechanisms with learning processes (see, for GBL, the LM-GM model by Arnab et al. [110], discussed in the context of bGBL by Abbott [52]) and d) identifying opportunities and tools for internal and external assessment. Addressing those issues is instrumental to achieve a better understanding of the factors that bGBL designers must consider for operating their choices, in turn providing better instruction to AI tools to assist on bGBL design. The criticalities that emerge from our analysis can be at least in part overcome by developing more operative instructional frameworks for bGBL. Despite those limitations, our study provides a strong indictment towards the potential of AI-assisted bGBL, supporting its unexplored potential for facilitating other difficult instructional steps in bGBL. This synergy will provide double benefits: in the first place, it will facilitate the adoption of bGBL in formal education settings, such as the school and the university, streamlining teacher training processes and instructional design tools. In turn, this will help the recognition and appreciation of games in education not just for their engagement and motivational properties, but as true learning environments for the development of significant knowledge, skills, and competencies [17, 49]. In the second place, it will facilitate the use of AI tools to assist instructional design by generating more consistent, context-sensitive, and accurate suggestions.

5. Conclusion

In the realm of educational technology and game-based learning (GBL), the use of generative AI tools to assist in the instructional design of board game-based learning (bGBL) and GBL on a broader scale represents a pioneering endeavor. As far as our current knowledge extends, this comprehensive analysis marks the first significant foray into exploring this synergy, specifically in the context of board games. This trailblazing work has profound implications for the future of educational design and technology integration, shedding light on uncharted territory; indeed, this study not only offers a novel perspective, but also exposes the hitherto underutilized potential of generative AI in shaping educational methodologies. It underscores that AI has the capacity to act as a transformative force in addressing the challenges that have historically hindered the effective implementation of GBL, including the creation of tailored content, scalability, and the need for personalized learning experiences. One of the study's implications is the need for the development of a comprehensive and integrated framework for instructional design in the context of board game-based activities. This framework would not only facilitate the seamless integration of AI but also offer a roadmap for designing educational experiences that harness the intrinsic appeal of board games for effective learning. In conclusion, the study's findings, while acknowledging its inherent limitations, showcases the tantalizing potential of AI and GBL to work in harmony, complementing each other's strengths. This synergy can pave the way for more effective, efficient, and personalized educational experiences.

Ethic statement

The authors declare no conflict of interest. The authors also declare that all experiments have been conducted in line with current ethical guidelines.

References

[1] A.I. Abdul Jabbar and P. Felicia, Gameplay engagement and learning in game-based learning: a systematic review, *Rev. Educ. Res.* **85** (2015), 740–779. doi:10.3102/0034654315577210

[2] M. Boeker, P. Andel, W. Vach and A. Frankenschmidt, Game-Based E-Learning Is More Effective than a Conventional Instructional Method: A Randomized Controlled Trial with Third-Year Medical Students, *PLOS ONE* **8** (2013), e82328. doi:10.1371/journal.pone.0082328

[3] J. Denner, S. Campe and L. Werner, Does computer game design and programming benefit children? A meta-synthesis of research *ACM Trans. Comput. Educ.* **19** (2019), 19:1–19:35. doi:10.1145/3277565

[4] V. Gashaj, L.C. Dapp, D. Trninic and C.M. Roebers, The effect of video games, exergames and board games on executive functions in kindergarten and 2nd grade: An explorative longitudinal study, *Trends Neurosci. Educ.* **21** (2021), 100162. doi:10.1016/j.tine.2021.100162

[5] T. Hainey, T.M. Connolly, E.A. Boyle, A. Wilson and A. Razak, A systematic literature review of games-based learning empirical evidence in primary education, *Comput. Educ.* **102** (2016), 202–223.

[6] S. Noda, K. Shirotsuki and M. Nakao, The effectiveness of intervention with board games: a systematic review, *Biopsychosoc. Med.* **13** (2019).

[7] J.L. Plass, B.D. Homer, E.O. Hayward, J. Frye, T.-T. Huang, M. Biles, M. Stein and K. Perlin, *The Effect of Learning Mechanics Design on Learning Outcomes in a Computer-Based Geometry Game*, E-Learn. Games Train. Educ. Health Sports, Springer, Berlin, Heidelberg, (2012), pp. 65–71.

[8] B. Yurdaarmağan, C.G. Melek, B. Merdenyan, O. Cikrikcili, Y.B. Salman and H. Cheng, The effects of digital game-based learning on performance and motivation for high school students (2015), http://arelarsiv.arel.edu.tr/xmlui/handle/20.500.12294/2044

[9] Y. Allsop, E. Yildirim and M. Screpanti, *Teachers' Beliefs About Game Based Learning: A Comparative Study of Pedagogy, Curriculum and Practice in Italy, Turkey and the UK* (2013). https://avesis.uludag.edu.tr/yayin/95fb7a0b-dfe0-4a85-afc4-22f45518c35d/teachers-beliefs-about-game-based-learning-a-comparative-study-of-pedagogy-curriculum-and-practice-in-italy-turkey-and-the-uk

[10] Y. Allsop and J. Jessel, Teachers' Experience and Reflections on Game-Based Learning in the Primary Classroom: Views from England and Italy, *Int. J. Game-Based Learn.* **5** (2015), 1–17.

[11] M. Kangas, A. Koskinen and L. Krokfors, A qualitative literature review of educational games in the classroom: the teacher's pedagogical activities, *Teach. Teach. Rev. Educ. Res.* **23** (2017), 451–470. doi:10.1080/13540602.2016.1206523

[12] D. Persico, M. Passarelli, F. Pozzi, J. Earp, F. Dagnino and F. Manganello, Meeting players where they are: Digital games and learning ecologies, *Br. J. Educ. Technol. Rev. Educ. Res.* **50** (2019).

[13] L.M. Takeuchi and S. Vaala, Level up Learning: A National Survey on Teaching with Digital Games, *Joan Ganz Cooney Center at Sesame Workshop* (2014). https://eric.ed.gov/?id=ED555585

[14] M. Cantoia, A. Clegg and A. Tinterri, Training Teachers to Design Game-Based Learning Activities: Evidence from a Pilot Project, *Comput. Sch* (2023).

[15] M. Kangas, P. Siklander, J. Randolph and H. Ruokamo, Teachers' engagement and students' satisfaction with a playful learning environment, *Teach. Teach. Educ. Rev. Educ. Res.* **63** (2017), 274–284.

[16] F.F. Loperfido, A. Dipace and A. Scarinci, To Play Or Not To Play? A Case Study Of Teachers' Confidence And Perception With Regard To Digital Games At School *Ital. J. Educ. Technol.* **27** (2019), 121–138. doi:10.17471/2499-4324/1062.

[17] S. De Freitas, G. Rebolledo-Mendez, F. Liarokapis, G. Magoulas and A. Poulovassilis, Learning as immersive experiences: Using the four-dimensional framework for designing and evaluating immersive learning experiences in a virtual world, *Br. J. Educ. Technol.* **41** (2010), 69–85. doi:10.1111/j.1467-8535.2009.01024.x

[18] A. Foster and M. Shah, The ICCE Framework: Framing Learning Experiences Afforded by Games, *J. Educ. Comput. Res.* **51** (2015), 369–395. doi:10.2190/EC.51.4.a

[19] T. Hanghøj and C. Brund, *Teacher Roles And Positionings In Relation To Educational Games* (2011), 125–136. doi:10.2307/jj.608141.9

[20] T. Adiguzel, H. Kaya and F. Cansu, Revolutionizing education with AI: Exploring the transformative potential of ChatGPT, *Contemp. Educ. Technol.* **15** (2023), ep429. doi:10.30935/cedtech/13152

[21] D. Baidoo-Anu and L. Owusu Ansah, *Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning* (2023). doi:10.2139/ssrn.4337484

[22] A. Bozkurt, J. Xiao, S. Lambert, A. Pazurek, H. Crompton, S. Koseoglu, R. Farrow, M. Bond, C. Nerantzi, S. Honeychurch, M. Bali, J. Dron, K. Mir, B. Stewart, E. Costello, J. Mason, C.M. Stracke, E. Romero-Hall, A. Koutropoulos, C.M. Toquero, L. Singh, A. Tlili, K. Lee, M. Nichols, E. Ossiannilsson, M. Brown, V. Irvine, J.E. Raffaghelli, G. Santos-Hermosa, O. Farrell, T. Adam, Y.L. Thong, S. Sani-Bozkurt, R.C. Sharma, S. Hrastinski and P. Jandrić, Speculative futures on ChatGPT and generative artificial intelligence (AI): A collective reflection from the educational landscape, *Asian J. Distance Educ.* **18** (2023).

[23] S. Hopcan, E. Polat, M.E. Ozturk and L. Ozturk, Artificial intelligence in special education: a systematic review, *Interact. Learn. Environ.* **0** (2022), 1–19.

[24] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. HÜllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn and G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, *Learn. Individ. Differ.* **103** (2023), 102274.

[25] C.K. Lo, What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature, *Educ. Sci.* **13** (2023), 410.

[26] W. Xu and F. Ouyang, A systematic review of AI role in the educational system based on a proposed conceptual framework, *Educ. Inf. Technol* **27** (2022), 740–779. 4195–4223.

[27] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martínez-Maldonado, G. Chen, X. Li, Y. Jin and D. Gašević, Practical and ethical challenges of large language models in education: A systematic scoping review, *Br. J. Educ. Technol.* (2023).

AUTHOR COPY

[28] M. Zafari, J.S. Bazargani, A. Sadeghi-Niaraki and S.-M. Choi, Artificial Intelligence Applications in K-12 Education: A Systematic Literature Review, *IEEE Access.* **10** (2022), 61905–61921.

[29] A. Bozkurt and R.C. Sharma, Generative AI and Prompt Engineering: The Art of Whispering to Let the Genie Out of the Algorithmic World, *Asian J. Distance Educ.* **18** (2023).

[30] M. Farrokhnia, S.K. Banihashem, O. Noroozi and A. Wals, A SWOT analysis of ChatGPT: Implications for educational practice and research, *Innov. Educ. Teach. Int.* **0** (2023), 1–15.

[31] P. Sridhar, A. Doyle, A. Agarwal, C. Bogart, J. Savelka and M. Sakr, *Harnessing LLMs in Curricular Design: Using GPT-4 to Support Authoring of Learning Objectives* (2023), doi:10.48550/arXiv.2306.17459

[32] N.H.O. Ali, R. Jamian and M.F. Yasak, Encouraging Classroom Learning through Game-Based Learning (GBL) Approach, *Prog. Eng. Appl. Technol.* **2** (2021), 787–798.

[33] J.L. Plass, R.E. Mayer and B.D. Homer, *Handbook of Game-Based Learning,*, MIT Press, Cambridge, MA, USA, 2020.

[34] Y.-R. Shi and J.-L. Shih, Game factors and game-based learning design model, *Int. J. Comput. Games Technol.* (2015).

[35] S. Tobias, J.D. Fletcher and A.P. Wind, Game-Based Learning, in: J.M. Spector, M.D. Merrill, J. Elen and M.J. Bishop (Eds.) *Handb. Res. Educ. Commun. Technol. Rev. Educ. Res.* **85** (2015), 740–779.

[36] C. Perrotta, G. Featherstone, H. Aston and E. Houghton, Game-Based Learning: Latest Evidence and Future Directions, *National Foundation for Educational Research,* (2013). https://research.monash.edu/en/publications/game-based-learning-latest-evidence-and-future-directions

[37] M. Qian and K.R. Clark, Game-based Learning and 21st century skills: A review of recent research, *Comput. Hum. Behav.* **63** (2016), 50–58.

[38] S. Ejsing-Duun and T. Hanghøj, *Design Thinking, Game Design and School Subjects: What is the Connection?* (2019). doi:10.34190/GBL.19.143

[39] T. Hanghøj, Game-Based Teaching: Practices, Roles and Pedagogies, *New Pedagog. Approaches Game Enhanc. Learn. Curric. Integr.* (2013), 81–101.

[40] K. Salen, R. Torres, L. Wolozin, R. Rufo-Tepper and A. Shapiro, *New Pedagog. Approaches Game Enhanc. Learn. Curric. Integr*, MIT Press, 2010.

[41] J. Hamari, D.J. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke and T. Edwards, Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning, *Comput. Hum. Behav* **54** (2016), 170–179.

[42] M. Hartt, H. Hosseini and M. Mostafapour, Game On: Exploring the Effectiveness of Game-based Learning, *Plan. Pract. Res.* **35** (2020), 589–604.

[43] P. Westwood, *Inclusive and Adaptive Teaching: Meeting the Challenge of Diversity in the Classroom*, 2nd ed., Routledge, London, 2018.

[44] D. Parlett, *The Oxford History of Board Games*, Oxford University Press, 1999.

[45] A. Cardinot and J. Fairfield, Game-Based Learning to Engage Students With Physics and Astronomy Using a Board Game, *Int. J. Game-Based Learn.* **9** (2019), 42–57.

[46] G. Calleja, *Unboxed: Board Game Experience and Design*, MIT Press, 2022.

[47] K. Oksanen, T. Lainema and R. Hämäläinen, Learning from Social Collaboration: A Paradigm Shift in Evaluating Game-Based Learning, *Handb. Res. Serious Games Educ. Appl.* (2017), 41–65.

[48] L. Vasconcelos, M. Sousa, F. Ferreira and J. Pinheiro, COLLABORATING: modern board games and collaboratories as a tool for capacity building, *SN Soc. Sci.* **2** (2022), 190.

[49] R.Y. Bayeck, *Examining Board Gameplay and Learning: A Multidisciplinary Review of Recent Research* (2020). doi:10.25384/SAGE.c.4943763

[50] M. Berland and V.R. Lee, Collaborative Strategic Board Games as a Site for Distributed Computational Thinking *Int. J. Game-Based Learn* **1** (2011), 65–81.

[51] P. Wonica, Learning to Evaluate Analog Games for Education, *Analog Game Studies*, (2017), 740–779. https://analoggamestudies.org/2015/05/evaluating-educational-goals-in-party-games/

[52] D. Abbott, *Modding Tabletop Games for Education*, in: M. Gentile, M. Allegra and H. Söbke (Eds.), Games Learn. Alliance, Springer International Publishing, Cham, 2019, pp. 318–329.

[53] M.J. Heron, P.H. Belford, H. Reid and M. Crabb, Meeple Centred Design: A Heuristic Toolkit for Evaluating the Accessibility of Tabletop Games, *Comput. Games J. Rev. Educ. Res.* **7** (2018), 97–114.

[54] M. Sousa, *Gamifying Serious Games: Modding Modern Board Games to Teach Game Potentials*, in: U. Dhar, J. Dubey, V. Dumblekar, S. Meijer and H. Lukosch (Eds.), Gaming Simul. Innov. Chall. Oppor., Springer International Publishing, Cham, 2022, 254–272.

[55] A.B. Heim and E.A. Holt, From Bored Games to Board Games: Student-Driven Game Design in the Virtual Classroom, *J. Microbiol. Biol. Educ.* **22** (2021).

[56] P. Parekh, E. Gee, K. Tran, E. Aguilera, L.E. Pérez Cortés, T. Kessner and S. Siyahhan, Board game design: an educational tool for understanding environmental issues, *Int. J. Sci. Educ. Rev. Educ. Res.* **43** (2021), 740–779.

[57] J.P. Zagal, J. Rick and I. Hsi, Collaborative games: Lessons learned from board games, *Simul. Gaming* **37** (2006), 24–40.

[58] E. Castronova and I. Knowles, A Model of Climate Policy Using Board Game Mechanics, *Int. J. Serious Games* **2** (2015).

[59] D. Abbott, Intentional Learning Design for Educational Games: A Workflow Supporting Novices and Experts, *Learn. User Exp. Res.* (2020).

[60] C.L. Weitze, Designing for Learning and Play – The Smiley Model as Framework, *IDA Interact. Des. Archit.* (2016), 52–75.

[61] M. Sousa, Mastering Modern Board Game Design to Build New Learning Experiences: the MBGTOTEACH Framework, Int. J. Games Soc. Impact. **1** (2023), 68–93.

[62] M. Andreoletti and A. Tinterri, *Apprendere con i giochi. Esperienze di progettazione ludica*, Carocci, Roma (2023).

[63] L.A. Annetta, The "I's" Have It: A Framework for Serious Educational Game Design, *Rev. Gen. Psychol.* **14** (2010), 105–113.

[64] J. Schell, *The Art of Game Design: A Book of Lenses*, CRC Press, 2019.

[65] J. Biggs, Enhancing Teaching through Constructive Alignment, *High. Educ.* **32** (1996), 347–364.

[66] A.R. Denham, Improving the Design of a Learning Game Through Intrinsic Integration and Playtesting, *Technol. Knowl. Learn. Rev. Educ. Res.* **21** (2016), 175–194.

AUTHOR COPY

[67] S. Nicholson, Making Gameplay Matter: Designing Modern Educational Tabletop Games, *Knowl. Quest.* **40** (2011), 60–65.

[68] D. Ifenthaler, D. Eseryel and X. Ge, *Assessment for Game-Based Learning*, in: D. Ifenthaler, D. Eseryel and X. Ge (Eds.), Assess. Game-Based Learn. Found. Innov. Perspect., Springer, New York, NY, 2012.

[69] N.B. Sardone, Modifying Board Games in Alignment with State Standards to Develop the Geographic Literacy of Elementary Level Learners, *Soc. Stud.* **0** (2022), 1–11.

[70] N.B. Sardone and R. Devlin-Scherer, Let the (Board) Games Begin: Creative Ways to Enhance Teaching and Learning, *Clear. House J. Educ. Strateg. Issues Ideas. Rev. Educ. Res.* **89** (2016), 215–22.

[71] C. van Esch and T. Wiggen, Can Your Students Save the World? Utilizing Pandemic, a Cooperative Board Game, to Teach Management, *Manag. Teach. Rev.* **5** (2020), 275–283.

[72] H. Desurvire, M. Caplan and J.A. Toth, *Using heuristics to evaluate the playability of games*, in: CHI 04 Ext. Abstr. Hum. Factors Comput. Syst., ACM, Vienna, 2004, pp. 1509–1512.

[73] M. Sasupilli, P. Bokil and R.M. Punekar, *Game Design Frameworks and Evaluating Techniques for Educational Games: A Review*, in: A. Chakrabarti (Ed.), Res. Des. Connect. World, Springer, Singapore, 2019, pp. 277–286.

[74] M. Tuomisto and M. Aksela, Design and evaluation framework for relevant chemistry-related educational card and board games, *LUMAT Int. J. Math Sci. Technol. Educ.* **3** (2015), 429–438.

[75] J. Hu, Teaching Evaluation System by use of Machine Learning and Artificial Intelligence Methods, *Int. J. Emerg. Technol. Learn. IJET* **16** (2021), 87–101.

[76] A.A. Wagan, A.A. Khan, Y.-L. Chen, P.L. Yee, J. Yang and A.A. Laghari, Artificial Intelligence-Enabled Game-Based Learning and Quality of Experience: A Novel and Secure Framework (B-AIQoE), *Sustainability* **15** (2023), 5362.

[77] Y.Y. Dyulicheva and A.O. Glazieva, *Game based learning with artificial intelligence and immersive technologies: an overview*, in: A.E. Kiv, S.O. Semerikov, V.N. Soloviev and A.M. Striuk (Eds.), Proc. 4th Workshop Young Sci. Comput. Sci. Softw. Eng. CSSESW 2021, CEUR, Virtual Event, Kryvyi Rih, 2021, pp. 146–159.

[78] N. Henderson, J. Rowe, L. Paquette, R.S. Baker and J. Lester, Improving affect detection in game-based learning with multimodal data fusion, *Artif. Intell. Educ.- 21st Int. Conf. AIED 2020 Proc. Part I* (2020), 228–239.

[79] Z. Zhan, Y. Tong, X. Lan and B. Zhong, A systematic literature review of game-based learning in Artificial Intelligence education, *Interact. Learn. Environ.* **0** (2022), 1–22.

[80] E. Southgate, K. Blackmore, S. Pieschl, S. Grimes, J. McGuire and K. Smithers, *Artificial intelligence and emerging technologies in schools: Research report*, University of Newcastle, Newcastle, NSW, 2019.

[81] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han and Y. Tang, A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development, *IEEECAA J. Autom. Sin.* **10** (2023), 1122–1136.

[82] M. Bearman, J. Ryan and R. Ajjawi, Discourses of artificial intelligence in higher education: a critical literature review, *High. Educ.* **86** (2023), 369–385.

[83] I. Celik, Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education, *Comput. Hum. Behav.* **138** (2023), 107468.

[84] A. Zirar, Exploring the impact of language models, such as ChatGPT, on student learning and assessment *Rev. Educ.* **11** (2023).

[85] L.K. Ch'ng, How AI Makes its Mark on Instructional Design, *Asian Journal of Distance Education* **18**(2) (2023), 32–41. doi:10.5281/ZENODO.8188576

[86] M. Molenda, In search of the elusive ADDIE model, *Perform. Improv.* **42** (2003), 34–36.

[87] A. Tinterri, M. di Padova, F. Palladino, G. Vignoli and A. Dipace, AI in board Game-Based Learning, *Proc. First Int. Workshop High-Perform. Artif. Intell. Syst. Educ.*, (2024). https://ceur-ws.org/Vol-3605/12.pdf

[88] S. Cuomo, G. Biagini and M. Ranieri, Artificial Intelligence Literacy, che cos'è e come promuoverla. Dall'analisi della letteratura ad una proposta di Framework, *Media Educ.* (2022).

[89] D.T.K. Ng, J.K.L. Leung, S.K.W. Chu and M.S. Qiao, Conceptualizing AI literacy: An exploratory review, *Comput. Educ. Artif. Intell.* **2** (2021), 100041.

[90] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocoń, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Miłkowski, M. Oleksy, M. Piasecki, Ł. Radliński, K. Wojtasik, S. Woźniak and P. Kazienko, ChatGPT: Jack of all trades, master of none, *Inf. Fusion.* **99** (2023).

[91] European Parliament, REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), (2024).

[92] X. Lu, S. Fan, J. Houghton, L. Wang and X. Wang, *ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions*, (2023), 1–18.

[93] T.C. Chen, E. Kaminski, L. Koduri, A. Singer, J. Singer, M. Couldwell, J. Delashaw, A. Dumont and A. Wang, Chat GPT as a Neuro-Score Calculator: Analysis of a Large Language Model's Performance on Various Neurological Exam Grading Scales, *World Neurosurg* **179** (2023), 1–18.

[94] G.A. Guerra, H. Hofmann, S. Sobhani, G. Hofmann, D. Gomez, D. Soroudi, B.S. Hopkins, J. Dallas, D.J. Pangal, S. Cheok, V.N. Nguyen, W.J. Mack and G. Zada, GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students and Neurosurgery Residents on Neurosurgery Written Board-Like Questions, *World Neurosurg* **179** (2023), e160–e165.

[95] H. Wang, W. Wu, Z. Dou, L. He and L. Yang, Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI, *Int. J. Med. Inf.* **177** (2023).

[96] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi and G. Neubig, Pre-train, Prompt and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, *ACM Comput. Surv.* **55** (2023), 195:1–195:35.

[97] E. Theophilou, C. Koyutürk, M. Yavari, S. Bursic, G. Donabauer, A. Telari, A. Testa, R. Boiano, D. Hernandez-Leo, M. Ruskov, D. Taibi, A. Gabbiadini and D. Ognibene *Learning to Prompt in the Classroom to Understand AI Limits: A Pilot Study*, in: R. Basili, D. Lembo, C. Limon-

AUTHOR COPY

gelli and A. Orlandini (Eds.), AIxIA 2023 – Adv. Artif. Intell., Springer Nature Switzerland, Cham, 2023, pp. 481–496.

[98] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith and D.C. Schmidt, *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*, (2023). http://arxiv.org/abs/2302.11382

[99] L. Liu, Analyzing the Text Contents Produced by ChatGPT: Prompts, Feature-Components in Responses and a Predictive Model, *J. Educ. Technol. Dev. Exch. JETDE.* **16** (2023), 49–70.

[100] P. Parrish, Pedagogical Models for Compelling Learning Experiences with Technology, *EDEN Conf. Proc.* (2020), 247–256.

[101] P. Lyon, *Design Education: Learning, Teaching and Researching Through Design*, Routledge, London, 2016.

[102] D.H. Jonassen and L. Rohrer-Murphy, Activity theory as a framework for designing constructivist learning environments, *Educ. Technol. Res. Dev.* **47** (1999), 61–79.

[103] S.P. Greenhalgh, M.J. Koehler and L.O. Boltz, The Fun of Its Parts: Design and Player Reception of Educational Board Games, *Contemp. Issues Technol. Teach. Educ.* **19** (2019), 469–497.

[104] A.P.O.S. Vermeeren, E. Lai-Chong Law, V. Roto, M. Obrist, J. Hoonhout and K. Väänänen-Vainio-Mattila, User Experience Evaluation Methods: Current State and Development Needs. *Proc. 6th Nord. Conf. Hum.-Comput. Interact. Extending Boundaries* (2010), 521–530.

[105] S. Kraiczy, Á. Cseh and D. Manlove, On weakly and strongly popular rankings, *Discrete Appl. Math.* **340** (2023), 134–152.

[106] M.G. Kendall, A New Measure of Rank Correlation *Biometrika* **30** (1938), 81–93.

[107] R. Newson, Parameters behind "Nonparametric" Statistics: Kendall's tau, Somers' D and Median Differences, *Stata J.* **2** (2002), 45–64.

[108] J. Mazurek, Evaluation of ranking similarity in ordinal ranking problems, *Acta Acad. Karviniensia, Rev. Educ. Res.* **11** (2011), 119–128.

[109] J.G. Kemeny, Mathematics without Numbers, *Daedalus* **88** (1959), 577–591.

[110] S. Arnab, T. Lim, M.B. Carvalho, F. Bellotti, S. de Freitas, S. Louchart, N. Suttie, R. Berta and A. De Gloria, Mapping learning and game mechanics for serious games analysis, *Br. J. Educ. Technol.* **46** (2015), 391–411.

AUTHOR COPY