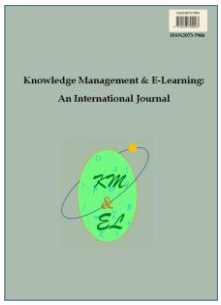
Knowledge Management & E-Learning, Vol.16, No.3. Sep 2024

A systematic review of game-based assessment in education in the past decade

Fan Su
Shanghai Normal University, Shanghai, China
Di Zou
The Hong Kong Polytechnic University, Hong Kong



Knowledge Management & E-Learning: An International Journal (KM&EL)
ISSN 2073-7904

Recommended citation:
Su, F., & Zou, D. (2024). A systematic review of game-based assessment in education in the past decade. *Knowledge Management & E-Learning, 16*(3), 451–476. https://doi.org/10.34105/j.kmel.2024.16.021

Knowledge Management & E-Learning, 16(3), 451–476

A systematic review of game-based assessment in education in the past decade

Fan Su

School of Education
Shanghai Normal University, Shanghai, China
E-mail: sufan0217@shnu.edu.cn

Di Zou*

Department of English and Communication
The Hong Kong Polytechnic University, Hong Kong
E-mail: dizoudaisy@gmail.com

*Corresponding author

Abstract: Educational games, prevalent in contemporary settings, leverage game-based learning (GBL) to actively engage and enhance learners' knowledge and skill acquisition through captivating in-game learning activities. To assess the effectiveness of GBL, game-based assessment (GBA) has emerged. GBA employs gameplay for learners to attain educational objectives while capturing data for analyzing their in-game competencies. The integration of gameplay in assessments has garnered increasing attention to GBA. This review of 21 studies provides an overview of the application of GBA in the field of education, covering aspects of the publication nature, theoretical frameworks, game types, in-game assessment details, and the assessed subjects and knowledge. Key findings include: (1) the annual number of publications fluctuates; a majority of studies originate from the USA; (2) the supporting theory tends to be unitary, with the evidence-centered design model cited the most; (3) simulation, immersive, and video games with rich game elements are applied most as assessment tools, and in-game assessment details are associated with game features; and (4) GBA is predominantly used in physics education. These findings indicate that applying GBA in education is promising with solid theoretical support, and practitioners are suggested to design GBA according to their educational needs and game features.

Keywords: Game-based assessment; Games; Education; Systematic review

Biographical notes: Fan Su received EdD degree at The Education University of Hong Kong and she works in Shanghai Normal University. Her research interests include second language writing, web-based collaborative writing, and technology-enhanced language learning. She has published research articles in international journals, such as Computer Assisted Language Learning, SAGE Open, International Journal of Mobile Learning and Organisation, International Review of Applied Linguistics in Language Teaching, and Technology, Knowledge and Learning.

Di Zou is an Associate Professor at the Hong Kong Polytechnic University in Hong Kong. Her research interests include second-language acquisition, technology-enhanced language learning, game-based language learning and flipped classroom. She has published more than 100 research papers in

international journals and books, including Computers & Education, Computer Assisted Language Learning, Language Teaching Research, and British Journal of Educational Technology.

1. Introduction

Educational games have captured researchers' attention as effective tools that engage and motivate learners (El Mawas et al., 2022), and game-based learning (GBL) has been shown to help learners improve academic achievement and motivation (Acquah & Katz, 2020; Hooshyar et al., 2021; Xie et al., 2021; Zou et al., 2021a; 2021b). The significance of game-based assessment (GBA) has become increasingly evident, driven by the rising interest in GBL and its effectiveness in evaluating learners' competencies (Chen et al., 2020). GBL appeals learners to engage with intrinsically interesting in-game learning activities (Wang et al., 2022; Zhang et al., 2023), mostly designed with task-progress difficulty (Hooshyar et al., 2018) to support learning and skill development (Hooshyar et al., 2021). GBA utilizes gameplay for learners to achieve learning objectives while simultaneously capturing data to analyze their competencies (Lin et al., 2020). In this case, GBA provides flexibility in designing the best combinations of authenticity and measurement efficiency, possibly reducing learners' anxiety on examination performance (Wang et al., 2022). Both GBL and GBA incorporate games into the learning process, with GBL emphasizing task completion, while GBA is capable of "building complex learning and assessment contexts into assessments which serve the purposes of both assessments of and for learning" (Chen et al., 2020, p. 482). GBA incorporates interactive tasks, gamified elements, and continuous recording of learners' cognitive processes and responses, serving to assess various dimensions of student learning. The integration of gameplay in assessments provides several advantages over conventional methods, such as storing real-time data, utilizing game logs for competency analysis, and establishing an engaging evaluative learning environment (Song & Sparks, 2019a). These features indicate that GBA utilizes empirical evidence and data-driven insights to measure learners' competencies and their learning and decision-making processes (Shute et al., 2007). Educational games typically empower learners to control their own learning, seamlessly evaluating their performance—a concept aligned with in-game stealth assessment, capable of reducing test anxiety without compromising reliability and validity (Shute, 2011). Thus, GBA lends itself well as a powerful tool in education assessment (Kim & Ifenthaler, 2019).

Several representative studies have illustrated the landscape of GBA. In a systematic review, Zhu et al. (2023) found that GBA was conducted across different educational levels, but the number of participants was small, with a primary focus on investigating their knowledge. Moreover, they observed that educational games were the most popular genre for implementing in-game assessments. Gomez et al. (2022) conducted a comprehensive review, revealing GBA's widespread application in K-16 education, the medical sector, and the workforce. Many studies focused on assessing GBA's suitability for educational environments. Kim and Ifenthaler (2019) highlighted the prevalence of the evidence-centered theory (ECD) as an assessment design framework within GBA studies and underscored the importance of analyzing learners' in-game actions to understand their learning processes. However, handling learners' affective states, processing in-game data, and delivering in-game feedback posed substantial challenges. In a targeted review, Bellotti et al. (2013) provided insights into assessments in specific serious games such as Immune Attack and SimVenture's

Business games. They emphasized serious games' advantages of seamlessly tracking players' movements and decisions during gameplay and the effective integration of pedagogy and games. Landers and Sanchez (2022), focusing on employee selection, emphasized the need to understand the complexities of GBA design, including algorithmic systems, players' emotional experiences, and the necessity to ensure GBA's stable assessment traits while aligning evaluation standards with the test claims.

Previous comprehensive reviews have extensively outlined the fields, theories, advances, and challenges associated with GBA. However, there is a noticeable lack of details on GBA implementation, particularly regarding assessments during gameplay and the types of games employed for in-game assessment. These gaps leave practitioners uncertain about suitable games for designing and administering assessments. Existing reviews, whether focused on GBA in a broad context or discussing a particular game type, lack an education-oriented perspective. Thus, an education-focused systematic review of GBA's application appears timely to provide insights for practitioners on harnessing games for assessment. To help practitioners comprehend GBA in education, this review aspires to unveil the current publication trends (i.e., the prolific years, regions, and journals), identify the theoretical framework of GBA, illustrate how this framework guides in-game assessment designs. Further, we aim to enrich the literature by categorizing game types, detailing in-game assessments, and reporting the assessed domains. This can guide educators across disciplines to employ GBA compatible with their subjects and inform designers on optimizing GBA's internal design for its practical application in education. The research questions guiding our study include:

1) What is the nature of publications focusing on the application of GBA in education?
2) What are the theoretical frameworks of the studies applying GBA in education?
3) What types of games are employed for educational assessment purposes?
4) How is in-game assessment designed and implemented?
5) What subjects and knowledge are assessed by GBA?

2. Literature review

In the digital era, accurately evaluating of learners' competencies during the learning process necessitates new technologies in educational measurement (Shute et al., 2008). The application of games for educational assessment is acknowledged as an effective strategy due to their ability to engage and motivate learners. GBA seamlessly integrates game design and assessment activities, striking a balance between player engagement and rigorous evaluation (Kim & Shute, 2015). Unlike traditional psychometrics, GBA utilizes interactive tasks and game-like elements, probing deeper into learners' knowledge, skills, thought processes, and strategies by analyzing actions learners naturally generate while performing in-game tasks (Shute et al., 2013). A critical step of GBA involves utilizing in-game data streams to inform educational instructions and enhance learners' learning. This is facilitated through the application of evidence-centered design, which prioritizes consistency between the evidence gathered within the game and the underlying knowledge or skills being assessed (Shute et al., 2007). Such an approach enhances the validity and reliability of assessments within educational games.

To offer a comprehensive picture of GBA, five representative studies on this theme are reviewed. Zhu et al. (2023) systematically reviewed 50 GBA studies published

from 2011 to 2022. Their results showed that (1) the United States was the most productive region; (2) most studies recruited fewer than 200 participants; (3) computers were the primary platform for GBA implementation; (4) popular game genres included adventure, simulation, strategy, role-playing, educational, and puzzle types, with educational games being predominant, and self-developed games were commonly used for research purposes; (5) GBA content focused on discipline-specific knowledge, followed by cognitive ability, contemporary competences, and affective states; (6) formative assessment with process data and summative assessment using final scores were common methods; and (7) correlation analysis was frequently applied to verify the effectiveness of GBA. While Zhu et al.'s (2023) review uncovered GBA details, especially regarding game genres for in-game assessment, it lacked specific information on the design and implementation of each GBA type, such as the practicality and applicability of games and learners' activities during gameplay. This gap hinders researchers and practitioners in selecting suitable games for in-game learning and assessment. To address this issue, a focused GBA review should explore in-game assessment details for various game genres.

Gomez et al. (2022) recently conducted a review of GBA studies up to January 2021, noting a growing interest in the research community since 2013. Their review highlighted GBA's prevalence in K-16 education, medicine, and the workforce, particularly within STEM and humanities domains. Analytical methods such as descriptive analysis, correlations, visualizations, and machine learning were most commonly adopted to examine the link between learners' behaviors and assessment outcomes. Limitations in GBA application included the need for reliable data analytic methods, recruitment of large samples, game validation, and alignment of game design with targeted constructs. While offering a broad understanding of GBA's application and limitations, Gomez et al.'s (2022) review lacked detailed insights into game types and in-game assessment particulars, which were crucial for designing GBA. Thus, a further review addresses these specifics is warranted.

Kim and Ifenthaler (2019) conducted a comprehensive analysis of GBA's trajectory from 2009 to 2019. Their findings revealed that ECD served as the primary guiding framework, with a recommendation for game designers to utilize automated assessment systems for formative and summative feedback. They advocated for analytics-driven GBA to assess learners' psychological states, providing real-time data on in-game behaviors, thereby transparentizing engagement and learning processes. Despite these advances, challenges in implementing GBA were acknowledge, including (1) balancing understanding of learners' in-game learning progression with addressing their affective, behavioral, and cognitive states; (2) employing advanced methodologies and technologies for internal supports like personalized feedback; (3) enhancing assessment literacy among stakeholders for data handling; and (4) utilizing ECD to shape game design around target competencies. While their review presented GBA's supporting theory, challenges, and advances, they did not delve into game types and in-game assessment details. This again indicates the need for a review specifically focused on these two aspects.

Focusing on specific games, Bellotti et al. (2013) provided an overview of assessment in serious games, emphasizing competence and skill assessment methods. They illustrated cases showcasing the proficiency of serious games-based assessment in tracking player movements and decisions during gameplay. This approach seamlessly integrated pedagogy and games, enabling immediate feedback and user adaptivity. Although this review offered a preliminary understanding of GBA and its benefits and problems, it lacked depth on the mechanics of designing and implementing the

assessment in these games. The review also highlighted topics such as assessing a game's effectiveness, while how gamified elements were designed to evaluate various dimensions of pre-set knowledge remained uncovered.

Landers and Sanchez (2022) highlighted key strategies for designing GBA in employee selection and emphasized the importance of understanding the complexities of GBA design. This included considerations of algorithmic systems, player-mechanic interactions, players' emotional experiences, and the concurrent pursuit of assessment goals and game development. They also underscored the need for GBA to exhibit stable assessment traits and align evaluation standards with test claims. Although Landers and Sanchez (2022) offered practical design tips for GBA, they did not identify what games have been applied and how in-game assessments have been organized.

While the above reviews offer valuable insights, a common limitation is their omission of the analysis of game types applied for assessments and the specifics of in-game assessment. This absence makes it challenging for practitioners to determine available and practical games for integrating particular assessments, clarify the implementation of in-game assessment details, and design needs-driven GBA. A comprehensive understanding of these aspects helps practitioners decipher assessment measures that evaluate learners' actions during gameplay and make informed choices about games tailored for purposeful assessment design and implementation. Furthermore, while many studies have reviewed GBL in educational contexts (e.g., Su et al., 2021), the exploration of GBA in education remains limited. This is a crucial area to explore, as game logs provide a novel avenue for assessing learners' learning processes and task completion steps leading to their final accomplishments. To address these shortcomings, our review focuses on GBA in education, specifically delving into game types and in-game assessment specifics, to guide educators in utilizing GBA in their subject teaching and informing future designers which aspects of GBA can be optimized for advancing its practicality in education.

3. Method

This review applied the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) coding scheme recommendation to select articles for inclusion (Moher et al., 2009). Fig. 1 presents a PRISMA-guided flowchart showing our systematic review process, including identification, screening, eligibility and inclusion.

3.1. Identification

We retrieved articles from Social Science Citation Index (SSCI) journals because they collect high-quality articles subjected to rigorous peer-review processes, ensuring reliability and authority. At the earlier stage, our review included book chapters and non-SSCI journals. However, upon examining these papers, we found that many did not satisfy our review standards due to lack of important details, particularly in elaborating how assessments were designed and implemented in games. For example, a non-SSCI study claimed to investigate GBA in academic writing for preservice teachers but lacked reporting on the game applied and how learners' in-game performances were collected for analysis and evaluation. Since the aim of our review is to examine game genres and in-game assessments, prerequisites for the selected articles include providing explicit information about game features and assessment details embedded in games. Otherwise,

the research questions cannot be adequately addressed. Therefore, we decided to limit our selections to SSCI articles for maintaining the standards set for this review.

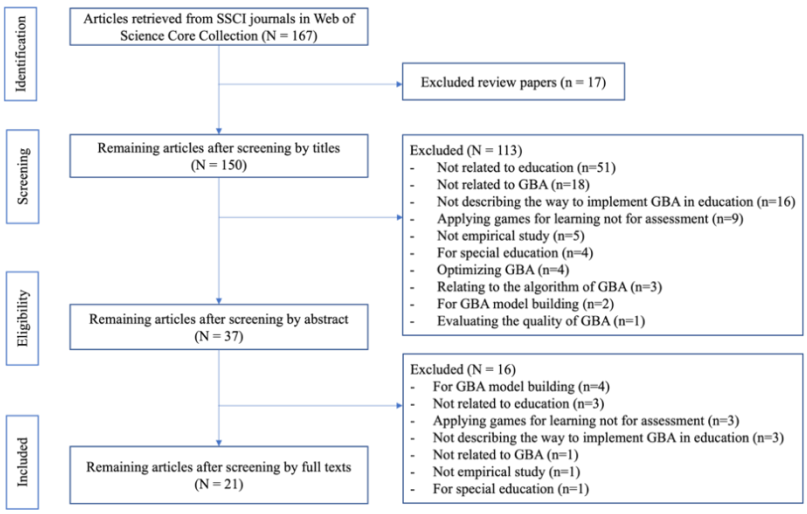


Fig. 1. Data screening procedures

Following the approach outlined by Gomez et al. (2022), we conducted a comprehensive search using combined keywords: "game-based assessment*," "game-related assessment*," "game-like assessment*," "stealth assessment*," "predictive learner modelling," "predictive player modeling," "knowledge tracing," with OR between them, and included the keyword "education." This search was performed through a paper's title, abstract and keywords to locate as many relevant studies as possible. Stealth assessment was used because it is seamlessly woven directly into the gameplay from which learner performance data are continually collected and inferences about learners' competencies are made (Kim & Ifenthaler, 2019; Shute, 2011). The time span was set as all years. Till the search date January 24, 2024, we have retrieved 167 English-written articles pertaining to GBA.

3.2. Inclusion and exclusion criteria

All 167 articles were checked based on inclusion and exclusion criteria. The inclusion criteria include (1) being empirical or experimental studies, (2) detailing the game features, and (3) elaborating how in-game assessments were designed and implemented.

Fig. 1 illustrates the application of exclusion criteria to remove irrelevant articles. Initially, 17 review papers were excluded based on their titles. Then, ten constraints were applied to remove irrelevant articles based on their abstracts, including (1) not relating to education, e.g., Hong and Liu (2022) incorporated GBA into the national identity investigation; (2) not relating to GBA, e.g., Flynn et al. (2021) utilized natural language process techniques to analyze the cohesion of readers' constructed responses; (3) not describing how GBA is implemented in education, e.g., Courtney and Graham (2019) investigated young learners' perceptions of digital GBA without elaborating on details; (4) applying games for learning not for assessment, e.g., Hooshyar et al. (2021) focused on the process and outcomes of GBL while not elaborating on GBA details; (5) non-empirical studies, e.g., Groff (2018) discussed the potentials of game-based environments for integrated, immersive learning data; (6) aiming for special education, e.g., Jobbágy et

al. (2016) developed assessment for children with birth injuries; (7) optimizing GBA, e.g., Georgiadis et al. (2021) proposed a generic stealth assessment software tool to improve the robustness of stealth assessment; (8) building the GBA model, e.g., Peters et al. (2021) constructed and validated a game-based intelligence assessment in Minecraft; (9) describing the algorithms of GBA, e.g., Levy (2019) explained a dynamic Bayesian network modelling approach for an educational video game; and (10) evaluating the quality of GBA; e.g., Hummel et al. (2017) discussed the way to assure the quality of GBA instead of incorporating it into educational contexts.

Implementing these exclusion criteria allowed our review focus specifically on studies addressing the application of GBA in educational settings. This approach facilitated a comprehensive investigation into the practical facets of GBA, grounded in empirical evidence and findings. The criteria effectively steered the focus away from delving into GBA for general purposes, discussions related to modelling, technical intricacies, and quality assessment. Consequently, the selected articles were directly pertinent to the precise dimensions of GBA in education targeted by the current review. Following full-text screening, 16 articles were further removed according to the above criteria, resulting in a final selection of 21 articles.

During the article screening process, we identified several studies centering around stealth assessment by Shute and her colleagues. However, these studies were non-empirical and lacked crucial information regarding the implementation of stealth assessments among participants and their experiences. The absence of such details hinder us to identify in-game assessment specifics, so the following articles were excluded: (1) Shute (2011)—published as a book chapter, focusing on elaborating the definitions and importance of stealth assessment as well as examples of stealth assessment systems; (2) Hansen et al. (2010) and Shute et al. (2007)—targeted individuals with special needs; (3) Shute and Becker (2010)—a book; (4) Shute et al. (2010)—published as a book chapter, focusing on modelling key competencies and developing valid assessments embedded within an immersive game; (5) Shute et al. (2009)—published as conference proceedings, providing future visions of assessment and learning in 21th century; (6) Shute and Zapata-Rivera (2008)—a research report, describing the role of educational assessment in intelligent systems; and (7) Shute et al. (2013)-a research report, focusing on an approach for embedding assessments in immersive games and recent advances in assessment design; (8) Shute and Spector (2008)—an unpublished manuscript envisioning a stealth assessment engine-enhanced system that can be run within games, simulations, and virtual worlds from methodological and theoretical perspectives; (9) Shute and Underwood (2006)—introducing technologies to assessment design from theoretical through addressing validity, equity, and access concerns.

Subsequently, we scrutinized the quality of all included articles to ensure that they not only met the criteria for inclusion but also provided comprehensive information crucial for addressing our research questions. Our rigorous evaluation focused on key aspects of the research design, including the recruitment of learners to participate into game-based learning activities, the seamless integration of games into educational contexts, detailed explanations and descriptions of the games employed, and elaborations of how assessments were incorporated into the games. This exhaustive examination aimed to go beyond surface-level scrutiny, achieving a comprehensive understanding of how learners' knowledge and skills were assessed during the gameplay. This meticulous inspection guaranteed the overall relevance of our reviewed articles.

3.3. Article coding scheme

Guided by research questions, we developed a coding scheme using a combination of inductive and deductive strategies to analyze 21 included articles. Deductive coding utilized pre-existing categories from literature, and inductive coding was employed when existing categories did not align. In the inductive coding processes, a bottom-up coding approach was further applied to merge similar contents and present them in a tabulated summary, minimizing subjectivity and providing an objective synthesis of the included articles. The authors collaboratively analyzed five articles to establish the coding scheme, during which they reminded each other to record coding details as literally reported in the articles. Subsequently, they independently coded the remaining articles. The results were compared, and satisfactory inter-rater reliability (r = 0.94) was achieved, with the differences resolved via discussion. These procedures worked together to create the following categories:

1) The publication nature (i.e., the publication year, areas, and journals) were coded as the articles reported.

2) The supporting theories serve as conceptual foundations guiding each research endeavor. In our reporting, we adhered to a transparent approach by presenting the theories exactly as articulated in the articles rather than making interpretations. To systematically organize the theories, all relevant contents were extracted and coded, with similar elements grouped together and summarized in tables. This coding process helped the researchers achieve clear and objective presentation of theoretical foundations of the reviewed articles. Specifically, studies grounded in the same theory were grouped together, while in studies where no specific theoretical support was mentioned or theories were irrelevant to GBA, the article was categorized with a code of "not specified." For instance, Song et al. (2023), Chen et al. (2020), and Song and Sparks (2019a, 2019b) identified evidence-centered design as their theoretical foundation. Consequently, these studies were categorized under the category of "Evidence-centered design." This classification was based on the acknowledgment that evidence-centered design, as described by Shute et al. (2007), serves as a conceptual framework for integrating assessments into educational games. Additionally, Cutumisu, Turgeon, et al. (2019) illustrated the Toulmin model used for developing learners' critical thinking, which was irrelevant to GBA. Kiili and Ketamo (2018) articulated the theoretical framework of an assessment triangle, while its connections with in-game learning and assessment were not clarified, so their theory was regarded as not specified.

3) To identify game genres, we documented game names and features (e.g., the gaming elements incorporated to engage students). This information was then compared with game genres reported by Su et al. (2021) and Zou et al. (2021a). For example, the Seaball game in our review immersed students in a fictional scenario to enhance reasoning skills, resonating with the description of immersive games where students used in-game support to complete learning tasks in a fictional world (Su et al., 2021). So, the Seaball game was categorized as an immersive game. If case of discrepancies, we coded the game genres as reported in the respective articles. Using this approach, game genres were finaslized as simulation, immersive, video, board and puzzle games, and gamification.

4) In-game assessments were meticulously summarized. Similarities in the designs of the same game used across different studies were grouped. For example,

Bergey et al. (2015) and Nelson et al. (2014) applied the Scientopolis game, addressing the issue of sick sheep on a virtual farm. However, the two studies differed in their assessment approaches. In Bergey et al. (2015), students answered questions to check their understanding of the sheep problem through multiple-choice questions. Nelson et al. (2014) collected students' interactions with signal and non-signal objects as assessment sources. These differences were presented as reported.

5) The subjects and knowledge assessed in each article were recorded. For example, Shute and Rahimi (2021) evaluated learners' grasp of physical principles using the Physics Playground. The investigated subject was physics, and the assessed knowledge was physical principles.

4. Results

4.1. Publication nature

Illustrated in Fig. 2, empirical research on applying GBA in education was limited until 2012. Notably, this does not imply the absence of GBA studies before 2012. Representative studies on stealth assessments existed but were excluded due to various reasons, such as being published as book chapters, focusing on modeling key competencies, or targeting individuals with special needs (Shute, 2011; Shute et al., 2007; Shute & Zapata-Rivera, 2008), lacking essential information (i.e., in-game assessment details) for our review objectives.

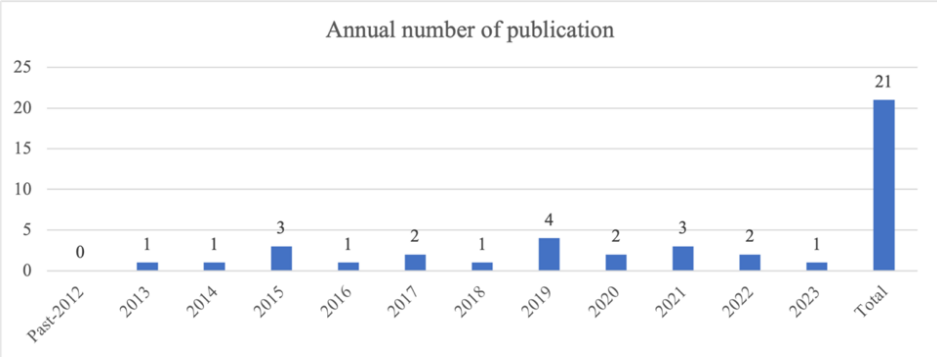


Fig. 2. Publication trend

Since 2013, scholars have increasingly focused on evaluating student learning within in-game settings. However, research interest has fluctuated over the following decade, possibly due to the time-intensive nature of designing in-game assessment procedures (Kim & Ifenthaler, 2019), and educators with a low level of technology acceptance may hesitate to design GBA.

Based on the affiliations of the first authors of the reviewed papers, we found that the included studies were conducted mainly in North America, Europe, and Asia (see Fig. 3). The USA was the most prolific area (N = 10, 48%), followed by Canada (N = 5, 24%) and Taiwan (N = 2, 10%) regions. The remaining four studies were distributed across Germany, Greece, Finland and mainland China, one in each area.

As shown in Fig. 4, the included articles were disseminated across 14 journals. Computers in Human Behavior and Computers & Education journals featured the highest number of studies applying GBA in education (three in each, 14%). Journal of Educational Computing Research, Journal of Computer Assisted Learning, and British Journal of Educational Technology each published two studies (10%). The remaining journals each included one publication.

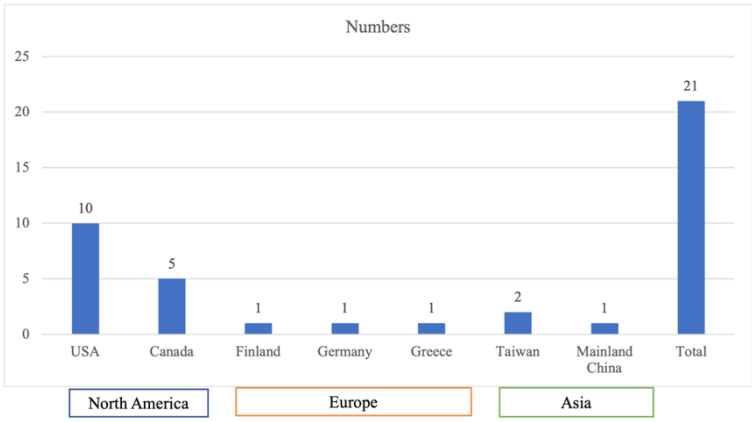


Fig. 3. Areas

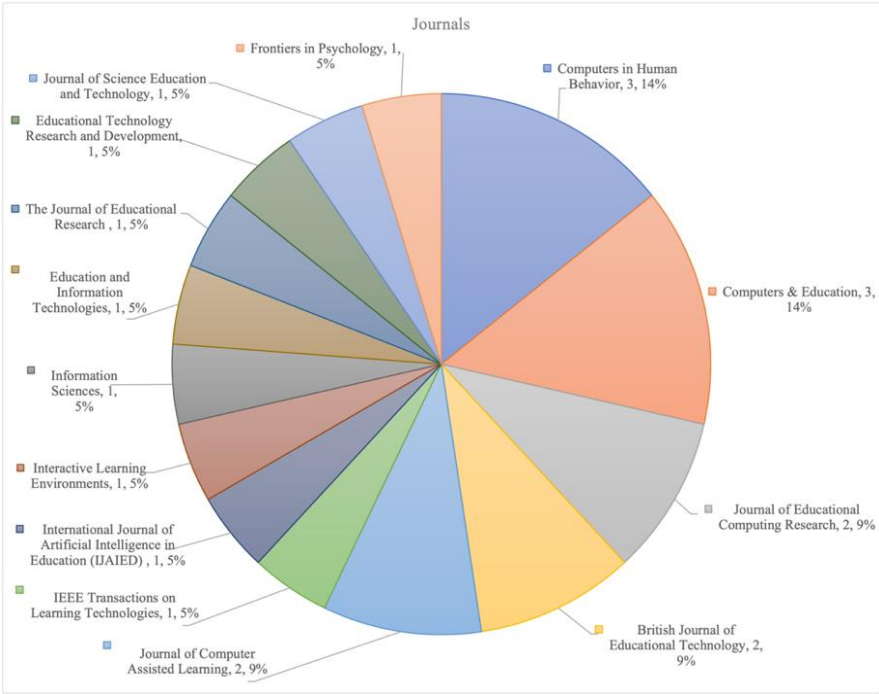


Fig. 4. Journals

4.2. Publication nature

Approximately 48% studies (N = 10) did not specify their supporting theories (see Table 1).

Table 1
Supporting theories

Theories	Studies
Not specified	Wang et al. (2022); Kim et al. (2023); Cutumisu, Turgeon, et al. (2019); Kiili and Ketamo (2018); Ninaus et al. (2017); Mavridis and Tsiatsos (2017); Bergey et al. (2015); Kim and Shute (2015); Tsai et al. (2015); Shute et al. (2013)
Conceptual design framework	
- Evidence-centered design	Song et al. (2023); Shute et al. (2021); Chen et al. (2020); Song and Sparks (2019a, 2019b); Shute et al. (2016)
Learner-centered assessment theories	
- Constructivist assessment & Choice-based assessment	Cutumisu and Schwartz (2021); Cutumisu, Chin, et al. (2019)
Cognitive theory	
- Guilford's theory of creativity	Shute and Rahimi (2021)
- Cognitive load theory	Nelson et al. (2014)
Unified theory of acceptance and use of technology	Lin et al. (2020)

The evidence-centered design theory was referred most (N = 6). ECD provides a conceptual design framework for crafting various educational assessment tests, emphasizing consistency between the evidence gathered and interpreted and the underlying knowledge addressed by assessments. In game-based environments, ECD guides researchers to collect evidence making valid claims about the level of students' competencies during gameplay (Shute et al., 2021). Researchers employing ECD specify the competences to be assessed, outline corresponding assessment tasks, and collect evidence (e.g., interactions, behaviors) during gameplay to infer students' mastery levels of the assessed competence (Shute et al., 2021; Shute et al., 2016; Song & Sparks, 2019a, 2019b; Song et al., 2023). For example, Shute et al. (2021) introduced the Physics Playground game, where learners manipulated a green ball to meet a red one using principles of physics; their in-game actions were analyzed to deduce their understanding of qualitative physics. In the simulation game Raging Skies, students used gamified tools to collect storm features, reflecting their skills to identify weather phenomenon (Chen et al., 2020). Song and Sparks (2019a, 2019b) and Song et al. (2023) designed a game Seaball enabling students to debate the junk food issue with virtual characters; evaluations of their arguments were related to the assessed reasoning skills. Shute et al. (2016) devised the Use Your Brainz game, where learners grew plants with diverse attributes to thwart zombies; their planting decisions were scrutinized via bayes nets to infer problem-solving abilities.

The category of learner-centered educational assessment includes constructivist assessment and choice-based assessment theories, both align with learner-centered paradigms (Cutumisu, Chin, et al., 2019; Cutumisu & Schwartz, 2021). Constructivist assessment emphasizes learners' active involvement in the learning process, with their knowledge constructed through meaningful interactions with the learning environment.

Choice-based assessment measures the learning choices students make during in-game action while solving challenges, connecting their learning in-game behaviors with real-world processes. Cutumisu and Schwartz (2021) and Cutumisu, Chin, et al. (2019) applied these theories in the Posterlet game where learners designed digital posters and selected feedback from virtual characters to decide whether to revise the posters. The feedback aligned with the graphic design principles: readability, crucial information, and spacing, offering learners learning chances while making revision choices.

The cognitive theory category encompasses Guilford's creativity and cognitive load theories. Guilford's creativity theory emphasizes the capacity to envision multiple solutions to a problem, incorporating creativity attributes through divergent thinking: flexibility (producing relevant problem-solving ideas), fluency (generating diverse pertinent ideas), originality (creating novel ideas), and elaboration (explicitly explaining an idea). Following this theory, Shute and Rahimi (2021) incorporated creativity assessments within the Physics Playground game, where students devised solutions to move a green ball towards a red one using gaming elements such as ramps and levers. Cognitive load theory addresses the mental load and effort imposed by a task on learners, assuming that learning requires active cognitive processing of storing information in long-term memory in the form of schemes that automates the process of retrieving information (Sweller et al., 1998). Building upon this theory, Nelson et al. (2014) inserted visual cues into the virtual world-based assessment of science inquiry to highlight the objects that students needed to interact with, increasing the likelihood they focused on instructive information in the virtual context and thereby reducing their cognitive load.

The unified theory of acceptance and use of technology (UTAUT) elucidates users' intentions to use technologies and subsequent usage behaviors, encompassing five key constructs: (1) performance expectancy (PE): students' perception of a technology's potential to enhance their achievements; (2) effort expectancy (EE): the ease of using a technology; (3) social influence (SI): others' opinions about the need to use a new technology and (4) facilitating conditions (FC): the support available to use a technology (Venkatesh et al., 2003). Utilizing the UTAUT framework, Lin et al. (2020) examined the moderating effects of computer-based and Kahoot!-based assessments on relationships between PE, EE, or SI and students' behavioral intention. They also explored how these assessment tools impacted students' learning performances in electronic commerce and changes in their perceptions of playfulness, behavioral intention, and use behavior associated with these assessment tools.

4.3. Game types

As shown in Fig. 5, simulation games constituted the highest usage (N = 7, 33%), followed by immersive (N = 6, 29%) and video games (N = 5, 24%). Gamification, board, and puzzle game was each only applied once (N = 1, 5%).

Simulation games replicate real life-activities to facilitate contextualized learning (Su et al., 2021). These games are typically integrated into curriculum-based scenarios, wherein learners employ virtual assistance to accomplish tasks (Bakan et al., 2022). Such immersive experiences provide students with opportunities to learn and subsequently apply strategies to address real-world challenges, leading to better decisions in their personal life (Cutumisu & Schwartz, 2021). In our review, examples of simulation games included Posterlet (Cutumisu, Chin, et al., 2019; Cutumisu & Schwartz, 2021; Cutumisu, Turgeon, et al., 2019), Raging Skies (Chen et al., 2020), Semideus (Kiili & Ketamo, 2018; Ninaus et al., 2017) and Noah Kingdom (Wang et al., 2022). These games simulate authentic situations within game environments where students controlled in-game

techniques to solve real-life problems, such as graphic design, weather phenomena, conceptual fraction knowledge, building critical thinking and evaluating arguments.

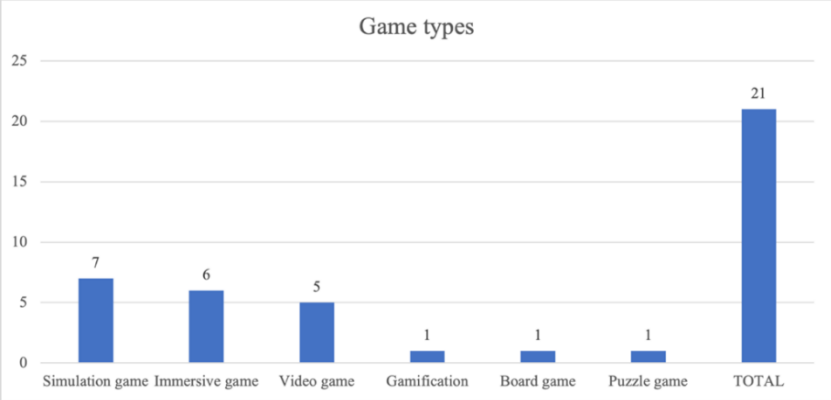


Fig. 5. Game types

Immersive games expose students to fictional realms, immersing them into overarching scenarios to complete tasks that evaluate their knowledge, which combines well-structured game contexts, seamless game flow, and lifelike realism (Su et al., 2021). Researchers developed Seaball (Song & Sparks, 2019a, 2019b; Song et al., 2023) for improving students' reasoning skills; Scientopolis (Bergey et al., 2015) and Save Science (Nelson et al., 2014) for improving students' science inquiry abilities; and a 3D treasure hunt game for facilitating students' multimedia system factual knowledge (Mavridis & Tsiatsos, 2017). These games created fictional social contexts (i.e., a broad program, a farm with sick sheep, and scientists on a remote island), wherein students tackled reality-like problems. They interacted with virtual characters, distinguishing their opinions or answering questions to receive feedback or obtain task clues. Before the end of these games, students were allowed to revisit these virtual characters to clarify their misunderstanding and reorganize useful information to complete subsequent tasks.

In video games, students apply dynamic game elements to achieve specific goals, with no fixed right or wrong answers judging their performances. During gameplay, students employ open-ended approaches to achieve predetermined goals, navigating possibilities by continually making choices and actions (Shute & Rahimi, 2021). Two video games were identified: Physics Playground (Kim & Shute, 2015; Shute et al., 2021; Shute & Rahimi, 2021; Shute et al., 2013) and Use Your Brainz (Shute et al., 2016), assessing students' conceptual understanding of physics knowledge and problem-solving skills respectively. In these games, students had unrestricted opportunities to devise optimal solutions for challenges, with their trajectories being recorded and further analyzed to evaluate their learning behaviors and their application of certain types of knowledge.

Gamification is not a game type but a design approach that integrates game elements into non-game contexts to make learning activities more game-like (Krath et al., 2021). In this review, gamification refers to learning designed with game mechanics and thinking to engage learners, motivate actions, and promote learning, which emphasizes either game elements (e.g., levels, points, badges, leaderboards, or certificate) or the process of gameful experiences in learning contexts (Su & Zou, 2022). Kahoot! was designed with game elements (e.g., graphics, progress bars, virtual rewards, leaderboards)

to create gameful learning experiences for facilitating students' electronic commerce knowledge (Lin et al., 2020).

Puzzle games guide students through a sequence of clues to enhance their problem-solving skills (Kim et al., 2023). In Shadowspect, students were given a collection of silhouettes from different views representing a figure they needed to create. To achieve this, students analyzed these silhouettes using their spatial reasoning skills and utilized given shapes (e.g., cubes, ramps) to establish the primitive figures.

Board games involve players to place, move, or remove pieces on a marked board with specific piece movement rules within a patterned layout (Noda et al., 2019). The Tic-tac-toe quiz (Tsai et al., 2015) in our review followed board game features, where students competed by placing pieces on a nine-square grid upon answering questions correctly, with the first to establish a row, column, or diagonal line declared as the winner.

4.4. In-game assessment details

4.4.1. Simulation game-based assessment

Simulation GBA progressively assessed learners' curriculum knowledge, and in-game assessment of their performances reflect their actions and responses in real life, guiding their decision-making in personal life (Cutumisu & Schwartz, 2021).

The simulation game *Posterlet* was frequently employed to assess students' graphic design knowledge and feedback behaviors (Cutumisu, Chin, et al., 2019; Cutumisu & Schwartz, 2021; Cutumisu, Turgeon, et al., 2019). In Posterlet, students participated in a simulated school environment, creating three posters for virtual booths by selecting text and images, arranging them on the canvas, and adjusting the appearance. Subsequently, virtual characters evaluated students' posters based on graphic design principles, and students chose confirmatory or critical feedback from each character for revision or submission. The game recorded the amount of critical feedback sought by students and the number of revisions made. Finally, students' poster scores and booth ticket sales were displayed, determined by the correct application of graphic design rules (Cutumisu, Chin, et al., 2019; Cutumisu & Schwartz, 2021). Differently, Cutumisu, Turgeon, et al. (2019) provided students with predetermined feedback in a specific order to gain an in-depth understanding of the processes unfolding during their interaction with critical feedback.

Raging Skies measured students' knowledge about weather phenomena (Chen et al., 2020) as they played a role of storm chasers controlling a gamified vehicle to gather information about real-time footage of storms. The difficulty of storm tasks was customized according to students' performance level. Students achieving at least 60% of in-game cash faced more challenging storm tasks, while those with less than 60% completed storms at the same difficulty level. In-game cash rewards for students included $500 for accurately identifying storm types; $100 ($50 for the second attempt) for correctly identifying storm elements, a $500 bonus for timely and accurate identifications; 100$ bonus for correctly identifying storm types and all storm elements. All students received reports summarizing their storm chasing accomplishments.

Semideus assessed students' grasp of conceptual fraction knowledge (Kiili & Ketamo, 2018; Ninaus et al., 2017). In the game, students directed the character *Semideus* to find the stolen gold coins encrypted in mathematical symbols and competed against a goblin to retrieve these coins by estimating the mathematic fraction. For inaccurate

estimation, students got 100-500 coins based on the degree of correctness and lose 15% units of energy. When students correctly compared the magnitudes of two fractions, their coin rewards were determined by response time. An incorrect comparison resulted in zero points and a 20% energy reduction. Level progression led to extra star awards, and a bonus was granted based on remaining energy after completing all tasks.

The *Noah Kingdom* game integrated interactive story scenes with progressive levels to foster critical thinking (Wang et al., 2022). Students participated in evidence collection, evaluation, querying, refutation, and argument matching to address embedded problems. In the background information collecting scene, students earned one point by identifying the correct evidence, while one point was deducted for two wrong attempts. In the dialogue-making scene, one point was awarded for a correct query and two points for a correct refute. In the decision-making scene, one point was granted for meeting one of the following criteria: claim consistency, positive evidence explanations, and strategy proposing according to the negative evidence.

4.4.2. Immersive game-based assessment

Immersive GBA yields rich data on learners' learning experiences when they completed the game tasks with different paces, worked on different items, and achieved different credits.

Song and Sparks (2019a, 2019b) and Song et al. (2023) applied the immersive game *Seaball*, prompting students to interact with virtual characters around the issue "whether junk food should be provided to school students." Students employed their reasoning skills across five activities: (1) interviewing virtual peers; (2) listening to a virtual speaker's opinions; (3) identifying arguments and supporting evidence; (4) making decisions and (5) creating food item classification criteria. In Song and Sparks (2019a), students earned full credit for correct answers to questions in activities (1), (3), and (4) on their initial attempt and revised their responses for partial credits after receiving feedback. For activities (2) and (5), students obtained full credits for correct answers and no credit for incorrect answers. In Song and Sparks (2019b), students gained points based on the identification and uses of appropriate evidence to support their classification of the junk food and their evaluation of virtual characters' arguments to identify fallacies. In Song et al. (2023), students needed to collect 10 out of 11 available opinions from virtual characters. Failure to recognize the character's opinions led to revisit the content. Upon successfully recognizing the opinion, students continued to select the evidence to support the character's opinion. After this, they classified the opinions into the ban or allow category. After classifying 10 opinions, a virtual character reviewed their classification and provided feedback, and then students revised the incorrect attempt.

Scientopolis and Save Science were the same game, requiring students to apply scientific inquiry skills to address sheep problems of dying (Bergey et al., 2015) and sickness (Nelson et al., 2014) on a virtual farm. In Bergey et al. (2015), students encountered sheep, farmers or residents. Sheep encounters prompted students to employ virtual tools for measurements such as leg, ear and body lengths, gender, and age. Residents' dialogues conveyed perceptions of the sheep problem, while farmers helped students check problem-solving progress. If the problem was solved, students answered questions pertaining to the sheep problem before proceeding to an end-of-module assessment featuring six multiple-choice questions on sheep features and their relations to health and adaptation on the farm. Nelson et al. (2014) designed visual and non-visual

cues when students solved the sheep problem. Visual cues employed red markers to emphasize objects, encouraging student interactions with these objects. In the non-signaling condition, all markers were removed, but the assessment was the same as in the visual signaling condition. In both conditions, students justified their conclusions using data about sheep features; student interaction frequencies with the objects, sheep and virtual character encounters, the number of measurements taken per sheep, and clipboard records were collected to gauge the assessment efficiency.

Mavridis and Tsiatsos (2017) applied a 3D *treasure hunt quiz* game, casting students as scientists discovering question objects with multiple-choice questions. Students had to answer questions correctly before the time ran out. Answers for each question and final scores were submitted to the instructor. The game concluded when a student completed all questions or when the time elapsed.

4.4.3. Video game-based assessment

Video GBA involves students using dynamic game elements to address problems, offering open-ended solutions without fixed answers (Shute et al., 2021).

Researchers applied Physics Playground to assess students' grasp of physics principles. In this game, students led a green ball to the red balloon by drawing machine-like devices, e.g., ramps, levers, pendulums, and springboards (Kim & Shute, 2015; Shute et al., 2021; Shute & Rahimi, 2021; Shute et al., 2013). Game difficulty varied across levels; each level was solved by specific physics parameters, and the solutions were assessed by a par value based on the minimum number of objects or attempts needed. Student achievements on levels earned them either a gold badge (using fewer than three objects to solve a level problem) or a silver badge (utilizing more than three objects) (Shute et al., 2021; Kim & Shute, 2015). During the gameplay, students' in-game performances, including correct and incorrect attempts in each level, time spent, restarts, and objects used, were recorded in log files. These files were later analyzed to deduce students' understanding of qualitative physics and their strengths and weaknesses in different physics aspects (Kim & Shute, 2015; Shute et al., 2013). Shute and Rahimi (2021) evaluated students' creativity by fluency (the number of objects drawn per solved and unsolved level), flexibility (the number of applicable agents attempted each level, standard deviation among agent frequencies of negative evidence, and consecutive use of incorrect agents), and originality (the differences between students' successful solutions and the most common one).

In the Use Your Brainz game, students strategically placed plants with diverse attributes to defend against virtual zombies (Shute et al., 2016). Their problem-solving skills and in-game interactions were assessed by bayesian nets that graphically presented the conditional dependencies between problem-solving variables (e.g., planning a solution pathway) and indicators (e.g., planting iceberg lettuce within the attack range). Bayesian nets for each game level were constructed because the indicators changed across levels varied in the difficulty.

4.4.4. Gamification-based assessment

In the gamification-based assessment, students responded to predetermined questions within a learning context, earning points or scores for correct answers. The total scores determined students' final ranks, and in-game performances were represented by badges indicating their learning achievements (Krath et al., 2021). In the Kahoot! gameplay,

students were given four color-coded answer panels. Once all students made their choices or the timer expired, the correct answer popped up. Their goal was to earn points for correct answers and speedy responses. Throughout the processes, students received no feedback on their scores and answer accuracy. The game ended with a leaderboard displaying the top five students (Lin et al., 2020).

4.4.5. Puzzle game-based assessment

Shadowspect, a puzzle game, assessed students' spatial reasoning skills (Kim et al., 2023). Using the given silhouettes from varied angles, students created a figure by selecting, composing, and adjusting a set of primitive shapes (e.g., cubes, cylinders and spheres). They then employed an embedded camera to capture silhouettes of the figure to check its proximity to the target. After students submitted their creations, the game proceeded with the evaluation and provided feedback.

4.4.6. Board game-based assessment

Tsai et al. (2015) applied the board game Tic-tac-toe quiz in a living technology course. In this game, students engaged in single-player or multi-player matches against the computer or peers, taking turns to place pieces on a nine-square grid when answering the pre-set questions correctly. At the end of the game, a score list showcased the top 10 performances and the top 10 players who obtained high correct answer ratios.

4.5. The assessed subjects and knowledge

Around 19% studies (N = 4) did not specify the specific subject fields where GBA was applied. The remaining studies predominantly used GBA in physics education (4 studies), followed by English language arts (3), mathematics (2), and science (2). Other applications covered various fields: geometry, electronic commerce, multimedia systems, and energy education (each once). Notably, a study by Cutumisu, Chin, et al. (2019) employed GBA to assess mixed learning achievements across English language arts, mathematics, and science.

Regarding the assessed knowledge, the evaluations focused on physical principles (N = 4) and argumentation skills (N = 3), followed by conceptual knowledge of fraction, feedback seeking behaviors and scientific inquiry (each twice). Other aspects: spatial reasoning skills, critical thinking, weather phenomenon, electronic commerce knowledge, graphic design principles, multimedia system factual knowledge, problem-solving skills, and energy knowledge, were each assessed once.

5. Discussion, implication and future direction

5.1. Discussion

Our findings, consistent with Kim and Shute (2015), indicate that GBAs remain in its nascent stages of development. Practitioners often hesitated to implement GBAs primarily due to a lack of knowledge on how to balance game design and assessment to maximize GBA effectiveness without compromising gamified features. However, the fact was that the engaging and motivating characteristics of games effectively immersed

students in learning tasks, which alleviates their perception of being under tests and consequently leading to their decreased test anxiety while increasing engagement (Kiili & Ketamo, 2018; Mavridis & Tsiatsos, 2017; Ninaus et al., 2017). Despite encountering challenging tasks, students remained engaged due to the availability of game-level support and feedback to get them out of learning stagnation, thereby sustaining their participation in the assessment activities and increasing the possibility of making learning progress. This sustained active participation in GBA tasks leads to rich in-game behaviors that can reveal students' cognitive processes and proficiencies in knowledge and skills. These aspects, collected through game logs, serve as evidence to predict students' learning competences, highlighting GBA's potential to predict various learning outcomes beyond assessing limited knowledge. This agrees with the notion that GBA can reflect students' authentic knowledge and skill in real-life situations (Cutumisu, Chin, et al., 2019). Additionally, applying games for educational assessment is promising because GBAs can predict students' future performance and lifelong learning. They capture and evaluate students' learning processes and behaviors in formal and informal settings via a short period of gameplay, effectively predicting a range of learning outcomes and reflecting the impact of in-game learning experiences on future problem-solving behaviors.

Echoing with Kim and Ifenthaler (2019), the ECD theory emerged as the most frequently cited, given its alignment with GBA's objective of collecting in-game evidence to make valid claims about players' competencies (Song & Sparks, 2019a; 2019b). Hence, researchers can consider integrating the ECD theory when crafting GBAs by identifying the competences needed to be assessed and predicted, deciding on the evidence required for competence prediction, designing in-game tasks to collect evidence, and finalizing task types, numbers and sequences. This approach enables the collection of valid evidence throughout students' gameplay, facilitating the inference of their genuine learning performances. Yet only two studies utilized multiple theories, i.e., the constructivist and choice-based assessment theories, aiming to enhance students' learning quality through communication and track their gameplay choices (Cutumisu & Schwartz, 2021; Cutumisu, Chin, et al., 2019). This highlights a limited variety of theoretical underpinnings for GBAs in education, despite its multifaceted nature containing games, assessments, and education. Diverse theoretical support is crucial for a more comprehensive understanding of GBAs.

Simulation, immersive, and video games have become prevalent assessment tools due to their rich game elements (e.g., feedback mechanism and user control). These features address students' evolving learning needs, enable seamless in-game assessments without disrupting engagement, and provide consecutive data for analyzing knowledge understanding and application over time. Notably, GBA's features largely correspond to the games' inherent characteristics. Simulation games, while offering problem-solving in curriculum-based scenarios, occasionally fall short due to the gap between simulated scenarios and real-life situations; students may react differently when facing real situations. Nevertheless, simulation games contributed to students' critical thinking and reasoning skills as they engage with problem-solving scenarios, which corroborates with Wang et al. (2022) who found simulation GBAs facilitative to students' critique abilities. Immersive games assess students' knowledge through diverse tasks in fictional worlds but struggle with incorporating traditional assessments like multiple-choice questions, given the personalized learning paces of students. Video games empower students with the freedom to manipulate game elements and explore various problem-solving options (Shute et al., 2021). Video GBAs lack a clear-cut right or wrong answer, so it offers subjective evaluation of decisions, differing from the objective nature of multiple-choice questions. Comparatively, gamification, puzzle, and board games exhibit less diversity in

their features, probably decreasing student engagement levels. Integrating progressively challenging course-related questions within these games, as suggested by Lin et al. (2020), can be a solution to enhance their effectiveness.

Continuous in-game assessments of players often combine player modeling methods with the analysis of specific player characteristics. In the realm of player modelling methods, evidence trace files, such as time-on-task, proficiency in specific skills or knowledge, decision-making sequences, are automatically recorded as time-stamped event logs to provide details generated by learners during the gameplay (Chen et al., 2020; Song & Sparks, 2019b; Song et al., 2023). For example, Song et al. (2023) collected the process data in the Seaball, where learners interviewed virtual characters on their opinions about junk food. These data encompassed students' moment-by-moment actions, including visits and interactions with different characters, time spent on each character, and the number of attempts. Analyzing these details provided insights into students' interaction sequences and difficult items that they felt hard to classify. This comprehensive analysis facilitates a deeper understanding of learners' argumentation skills. Analyzing these files helps identify challenging tasks within the game, recognize learner behavior patterns, and make inferences about their acquisitions of targeted skills. Using evidence trace files to conduct GBA aligns with the ECD theory——learners' generated behaviors and outputs in completing game tasks are collected to infer their mastery degrees of knowledge and skill competences. This approach minimizes construct-irrelevant variance between the skills measured, task designs, and learner data (Chen et al., 2020; Song & Sparks, 2019b). The data-driven analytics approach in player modeling is also a powerful strategy that involves collecting and analyzing learners' in-game performance data to adjust the game's difficulty level to better suit learners' current proficiency (e.g., Ninaus et al., 2017; Chen et al., 2020). This approach ensures that learners complete tasks aligning with their abilities, providing an accurate reflection of their knowledge and skills. It also helps identify learners' learning patterns that may not be immediately apparent through traditional outcome-based analysis, thereby improving the effectiveness, accuracy, and fairness of in-game assessments.

Games often provide immediate feedback and allow multi-round attempts to complete a task after receiving feedback (Ninaus et al., 2017; Tsai et al., 2015; Hooshyar et al., 2021). An important player characteristic modeled within the game is individuals' feedback-seeking preferences (i.e., confirmatory or critical), which correlates with their learning outcomes (Cutumisu & Schwartz, 2021). Individuals who select critical feedback tend to acquire more knowledge and skills than those selecting confirmatory one because confirmatory feedback emphasizes the knowledge that learners already know, whereas critical feedback provides new knowledge that can correct errors or misunderstandings in learners' cognitive structures (Cutumisu, Chin et al., 2019; etc.). Modelling this aspect allows accurate inferences about learners' skills, helping teachers provide scaffolding contingent on learners' needs (Song et al., 2023). Additionally, games usually model players' cognitive abilities, such as problem-solving skills (Shute et al., 2016), spatial reasoning (Kim et al., 2023), and creativity (Shute & Rahimi, 2021). Assessing these cognitive abilities enables researchers and educators to determine learners' current competency levels, as well as their strengths and weaknesses in specific cognitive aspects. Game designers can then adapt by incorporating solvable cognitive challenges at the edge of learners' abilities to maximize learning possibilities (Shute et al., 2021). Moreover, games can model learners' learning styles, such as visual or auditory preferences (Nelson et al., 2014). For example, Nelson et al. (2014) inserted visual cues into the virtual world-based assessment to highlight the objects that students needed to interact with. This personalization enhances the effectiveness of game content delivery,

aligning with learners' preferred instructional strategies and directing their attention to relevant information during learning or assessment.

5.2. Implications

The study highlights the importance of considering gender differences in GBA design, as male and female students have distinct ability beliefs in completing game tasks (Bergey et al., 2015; Kim & Shute, 2015). For example, females may show a lower need for challenge compared to males. Game designers are advised to conduct play-testing studies to understand how various design decisions, such as adaptive game levels, impact the behaviors of female and male students. The gathered data can then be utilized to construct assessment mechanics tailored to each gender.

Additionally, it is crucial to guide students on game features, as evidenced by Kim and Shute (2015), who discovered students' inappropriate utilization of the adaptivity function as they advanced to higher levels even without completing easier tasks. Therefore, practitioners should introduce game features and purposes before students' engagement in gameplay, enabling them to interact more effectively with the games. Practitioners also need to gather evidence on how and why students tend to prefer specific game features and how these features impact their in-game participation and GBA outcomes. In this way, students can take full advantage of game features to assist learning, and their in-game learning performance can be recorded utmost for assessment.

Furthermore, GBA aims to evaluate students' genuine learning performance, necessitating a balance between in-game assessments and external measures. When students deeply immersed into the gameplay, they may overlook the assessment components, thereby affecting the assessment accuracy. One feasible approach is integrating familiar assessment tasks such as blank filling into the game, which differs from written assessments only in its implementation medium, potentially maintaining the in-game assessment precision and student engagement.

Moreover, in-game feedback should be considered when designing GBA. Immediate feedback in games aids students in reflecting on learning processes. It is necessary to develop students' skills to evaluate the value of in-game feedback and seek feedback valences that lead to better learning outcomes (Cutumisu & Schwartz, 2021). Including both critical and confirmatory feedback within the game is recommended because these two types of feedback can provide equivalent information value and equal learning opportunities. Critical feedback prompts reflection on mistakes, especially when learning new skills, while confirmatory feedback reinforces positive behavior. Importantly, incorporating explanatory feedback for low achievers in GBA is needed because this can help them clarify the evaluative criteria. In contrast, explicit instructions can be given to high achievers to aid their differentiation of controversial and biased learning information. This is congruent with Song and Sparks (2019a), advocating tailored feedback and instructions based on students' learning proficiency to optimize GBA's practicality.

Last, three major challenges exit in designing GBA. The primary aspect is the careful consideration of the compatibility between GBA and the specific learning objectives of different courses. As indicated by Lin et al. (2020), GBA is more suitable for logical, arithmetic-based, procedure-based, or rule-based courses (e.g., math, computer science) demanding higher-level learning objectives than memory-based courses that require students less effort to master. This highlights the need for practitioners to recognize that GBA is not a one-size-fits-all strategy across diverse

educational contexts. The second consideration revolves around striking a balance between games and real tests (Mavridis & Tsiatsos, 2017). In games, students often become deeply immersed, potentially overlooking the test components and resulting in diminished learning performance. This challenge requires collaboration among game designers, practitioners, and researchers to find a balance between an implementation closely aligned with real tests and a design that is engaging but less effective as an evaluative method. The third aspect pertains to ensuring the assessment equality among students with varying levels of gaming experiences (Bergey et al., 2015). Notably, the design features of game environments can shape students' gaming experiences, which influences their game skills and competency beliefs, ultimately impacting their learning performance. Students with higher-level gaming experiences may exhibit superior in-game performance due to their adept utilization of gaming skills. Consequently, the design of game environments for educational purposes should consider the extent to which these environments offer equitable access for students with diverse experiences and perceptions as game players.

5.3. Future directions

We have identified potential future research directions for GBA from the reviewed articles. These include: (1) designing games with adaptive task levels based on students' in-game performance to capture their impromptu behaviors, serving as recourses to assess their gaming experiences (Shute et al., 2021); (2) investigating the discrepancies between students' behavioral intentions and their actual performances in assessment to uncover the underlying causes (Lin et al., 2020); (3) examining the factors that make GBA equitable for skilled and beginner gamers to cultivate a balanced gaming environment (Bergey et al., 2015); and (4) assessing students' cognitive abilities and their perceptions towards GBA to unveil its efficacy (Bergey et al., 2015; Mavridis & Tsiatsos, 2017; Shute et al., 2016).

6. Conclusion

Our review of 21 SSCI articles on GBA in education yielded several insights. The annual publication trends exhibit fluctuations, with a predominant origin of studies from the USA. Additionally, a majority of studies are based on the evidence-centered design model, indicating that the supporting theoretical framework tends to be unitary. Notably, assessment tools often feature rich gaming elements, such as simulation, immersive, and video games, with these elements influencing in-game assessment designs. Currently, GBA is predominantly applied in physics education. Analyses of these articles highlights the importance of considering learner genders, game feature guidance, and maintaining a balance between in-game assessments and external measures in GBA designs.

However, this review has certain limitations. Important themes such as psychometric analysis of GBA and player modeling methods and characteristics are omitted, which limits the depth of insights into in-game assessment traits and modeling mechanisms. So, these review themes need to be considered for covering multifarious aspects of GBA. Moreover, our focus solely on SSCI studies possibly exclude impactful studies from diverse sources, future reviews can broaden the article selection databases by including resources like ERIC and Scopus for a more holistic perspective on the subject.

Author Statement

The authors declare that there is no conflict of interest.

Acknowledgements

The authors would thank all reviewers and editors for their valuable comments and suggestions to this paper.

ORCID

Fan Su https://orcid.org/0000-0002-9327-673X

Di Zou https://orcid.org/0000-0001-8435-9739

References

References marked with an asterisk indicate studies included in the review.

Acquah, E. O., & Katz, H. T. (2020). Digital game-based L2 learning outcomes for primary through high-school students: A systematic literature review. *Computers & Education*, *143*: 103667. https://doi.org/10.1016/j.compedu.2019.103667

Bakan, U., Han, T., & Bakan, U. (2022). Learner perceptions and effectiveness of using a massively multiplayer online role-playing game to improve EFL communicative competence. *Knowledge Management & E-Learning*, *14*(3), 286–303. https://doi.org/10.34105/j.kmel.2022.14.016

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction*, *2013*(1): 136864. https://doi.org/10.1155/2013/136864

*Bergey, B. W., Ketelhut, D. J., Liang, S., Natarajan, U., & Karakus, M. (2015). Scientific inquiry self-efficacy and computer game self-efficacy as predictors and outcomes of middle school boys' and girls' performance in a science assessment in a virtual environment. *Journal of Science Education and Technology*, *24*(5), 696–708. https://doi.org/10.1007/s10956-015-9558-4

*Chen, F., Cui, Y., & Chu, M.-W. (2020). Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *International Journal of Artificial Intelligence in Education*, *30*(3), 481–503. https://doi.org/10.1007/s40593-020-00202-6

Courtney, L., & Graham, S. (2019). 'It's like having a test but in a fun way': Young learners' perceptions of a digital game-based assessment of early language learning. *Language Teaching for Young Learners*, *1*(2), 161–186. https://doi.org/10.1075/ltyl.18009.cou

*Cutumisu, M., Chin, D. B., & Schwartz, D. L. (2019). A digital game-based assessment of middle-school and college students' choices to seek critical feedback and to revise. *British Journal of Educational Technology*, *50*(6), 2977–3003. https://doi.org/10.1111/bjet.12796

*Cutumisu, M., & Schwartz, D. L. (2021). Feedback choices and their relations to learning are age-invariant starting in middle school: A secondary data analysis. *Computers & Education*, *171*: 104215. https://doi.org/10.1016/j.compedu.2021.104215

*Cutumisu, M., Turgeon, K.-L., Saiyera, T., Chuong, S., González Esparza, L. M.,

MacDonald, R., & Kokhan, V. (2019). Eye tracking the feedback assigned to undergraduate students in a digital assessment game. *Frontiers in Psychology*, *10*: 1931. https://doi.org/10.3389/fpsyg.2019.01931

El Mawas, N., Truchly, P., Podhradsky, P., Medvecky, M., & Muntean, C. H. (2022). Impact of game-based learning on STEM learning and motivation: Two case studies in Europe. *Knowledge Management & E-Learning*, *14*(4), 360–394. https://doi.org/10.34105/j.kmel.2022.14.020

Flynn, L. E., McNamara, D. S., McCarthy, K. S., Magliano, J. P., & Allen, L. K. (2021). The appearance of coherence: Using cohesive properties of readers' constructed responses to predict individual differences. *Revista Signos*, *54*(107), 1061–1088. https://doi.org/10.4067/S0718-09342021000301061

Georgiadis, K., Van Lankveld, G., Bahreini, K., & Westera, W. (2021). On the robustness of stealth assessment. *IEEE Transactions on Games*, *13*(2), 180–192. https://doi.org/10.1109/TG.2020.3020015

Gomez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2022). A systematic literature review of game-based assessment studies: Trends and challenges. *IEEE Transactions on Learning Technologies*, *16*(4), 500–515. https://doi.org/10.1109/TLT.2022.3226661

Groff, J. S. (2018). The potentials of game-based environments for integrated, immersive learning data. *European Journal of Education*, *53*(2), 188–201. https://doi.org/10.1111/ejed.12270

Hansen, E. G., Shute, V. J., & Landau, S. (2010). An assessment-for-learning system in mathematics for individuals with visual impairments. *Journal of Visual Impairment & Blindness*, *104*(5), 275–286. https://doi.org/10.1177/0145482X1010400503

Hong, X., & Liu, Q. (2022). Assessing young children's national identity through human-computer interaction: A game-based assessment task. *Frontiers in Psychology*, *13*: 956570. https://doi.org/10.3389/fpsyg.2022.956570

Hooshyar, D., Pedaste, M., Yang, Y., Malva, L., Hwang, G.-J., Wang, M., Lim, H., & Delev, D. (2021). From gaming to computational thinking: An adaptive educational computer game-based learning approach. *Journal of Educational Computing Research*, *59*(3), 383–409. https://doi.org/10.1177/0735633120965919

Hooshyar, D., Yousefi, M., Wang, M., & Lim, H. (2018). A data-driven procedural-content-generation approach for educational games. *Journal of Computer Assisted Learning*, *34*(6), 731–739. https://doi.org/10.1111/jcal.12280

Hummel, H. G. K., Joosten-ten Brinke, D., Nadolski, R. J., & Baartman, L. K. J. (2017). Content validity of game-based assessment: Case study of a serious game for ICT managers in training. *Technology, Pedagogy and Education*, *26*(2), 225–240. https://doi.org/10.1080/1475939X.2016.1192060

Jobbágy, Á., Schultheisz, J., Horváth, M., & Réfy Vraskóné, H. (2016). Development of an effective therapy and objective assessment for children with birth injuries. *International Journal of Rehabilitation Research*, *39*(4), 354–360. https://doi.org/10.1097/MRR.0000000000000179

*Kiili, K., & Ketamo, H. (2018). Evaluating cognitive and affective outcomes of a digital game-based math test. *IEEE Transactions on Learning Technologies*, *11*(2), 255–263. https://doi.org/10.1109/TLT.2017.2687458

Kim, Y. J., & Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-Based Assessment Revisited* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-030-15569-8_1

*Kim, Y. J., Knowles, M. A., Scianna, J., Lin, G., & Ruipérez-Valiente, J. A. (2023). Learning analytics application to examine validity and generalizability of game-based

assessment for spatial reasoning. *British Journal of Educational Technology*, *54*(1), 355–372. https://doi.org/10.1111/bjet.13286

*Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education*, *87*, 340–356. https://doi.org/10.1016/j.compedu.2015.07.009

Krath, J., Schürmann, L., & Von Korflesch, H. F. O. (2021). Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior*, *125*, 106963. https://doi.org/10.1016/j.chb.2021.106963

Landers, R. N., & Sanchez, D. R. (2022). Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. *International Journal of Selection and Assessment*, *30*(1), 1–13. https://doi.org/10.1111/ijsa.12376

Levy, R. (2019). Dynamic bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, *54*(6), 771–794. https://doi.org/10.1080/00273171.2019.1590794

*Lin, J.-W., Tsai, C.-W., & Hsu, C.-C. (2020). A comparison of computer-based and game-based formative assessments: A long-term experiment. *Interactive Learning Environments*, *31*(2), 938–954. https://doi.org/10.1080/10494820.2020.1815219

*Mavridis, A., & Tsiatsos, T. (2017). Game-based assessment: Investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning*, *33*(2), 137–150. https://doi.org/10.1111/jcal.12170

*Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-Analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

*Nelson, B. C., Kim, Y., Foshee, C., & Slack, K. (2014). Visual signalling in virtual world-based assessments: The SAVE Science project. *Information Sciences*, *264*, 32–40. https://doi.org/10.1016/j.ins.2013.09.011

*Ninaus, M., Kiili, K., McMullen, J., & Moeller, K. (2017). Assessing fraction knowledge by a digital game. *Computers in Human Behavior*, *70*, 197–206. https://doi.org/10.1016/j.chb.2017.01.004

Noda, S., Shirotsuki, K., & Nakao, M. (2019). The effectiveness of intervention with board games: A systematic review. *BioPsychoSocial Medicine*, *13*(1), 22. https://doi.org/10.1186/s13030-019-0164-1

Peters, H., Kyngdon, A., & Stillwell, D. (2021). Construction and validation of a game-based intelligence assessment in minecraft. *Computers in Human Behavior*, *119*, 106701. https://doi.org/10.1016/j.chb.2021.106701

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, *55*(2), 503–524.

Shute, V. J., & Becker, B. J. (2010). *Innovative assessment for the 21st century*. Springer.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). An assessment for learning system called aced: Designing for learning effectiveness and accessibility. *ETS Research Report Series*, *2007*(2): i-45. https://doi.org/10.1002/j.2333-8504.2007.tb02068.x

Shute, V. J., Masduki, I., Donmez, O., Dennen, V. P., Kim, Y.-J., Jeong, A. C., & Wang, C.-Y. (2010). Modeling, assessing, and supporting key competencies within game environments. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge* (pp. 281–309). Springer US. https://doi.org/10.1007/978-1-4419-5662-0_15

*Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*: 106647.

https://doi.org/10.1016/j.chb.2020.106647

Shute, V. J., & Spector, J. M. (2008). SCORM 2.0 white paper: Stealth assessment in virtual worlds. Unpublished manuscript. Retrieved from https://www.researchgate.net/profile/Valerie-Shute/publication/228654079_SCORM_20_white_paper_Stealth_assessment_in_virtual_worlds/links/0fcfd5092c3687ff03000000/SCORM-20-white-paper-Stealth-assessment-in-virtual-worlds.pdf

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2008). Monitoring and fostering learning through games and embedded assessments. *ETS Research Report Series, 2008*(2): i-32. https://doi.org/10.1002/j.2333-8504.2008.tb02155.x

*Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's playground. *The Journal of Educational Research*, *106*(6), 423–430. https://doi.org/10.1080/00220671.2013.832970

*Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Shute, V. J., & Zapata-Rivera, D. (2008). Educational assessment using intelligent systems. *ETS Research Report Series*, *2008*(2): i-15. https://doi.org/10.1002/j.2333-8504.2008.tb02154.x

Shute, V., Levy, R., Baker, R., Zapata, D., & Beck, J. (2009). Assessment and learning in intelligent educational systems: A peek into the future. In *Proceedings of AIED 2009: 14th International Conference on Artificial Intelligence in Education Workshops*. Retrieved from https://aribadernatal.com/docs/AIED2009-IEG-WorkshopProceedings-FINAL.pdf#page=107

*Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning*, *37*(1), 127–141. https://doi.org/10.1111/jcal.12473

Shute, V., & Underwood, J. (2006). Diagnostic assessment in mathematics problem solving. *Technology Instruction Cognition and Learning*, *3*(1/2), 151–166.

*Song, Y., & Sparks, J. R. (2019a). Measuring argumentation skills through a game-enhanced scenario-based assessment. *Journal of Educational Computing Research*, *56*(8), 1324–1344. https://doi.org/10.1177/0735633117740605

*Song, Y., & Sparks, J. R. (2019b). Building a game-enhanced formative assessment to gather evidence about middle school students' argumentation skills. *Educational Technology Research and Development*, *67*(5), 1175–1196. https://doi.org/10.1007/s11423-018-9637-3

*Song, Y., Zhu, M., & Sparks, J. R. (2023). Exploring the role of process data analysis in understanding student performance and interactive behavior in a game-based argument task. *Journal of Educational Computing Research*, *61*(5), 1096–1120. https://doi.org/10.1177/07356331221138734

Su, F., & Zou, D. (2022). Learning English with the mobile language learning application "Duolingo": The experiences of three working adults at different proficiency levels. *International Journal of Mobile Learning and Organisation*, *16*(4), 409–428. https://doi.org/10.1504/IJMLO.2022.125959

Su, F., Zou, D., Xie, H., & Wang, F. L. (2021). A comparative review of mobile and non-mobile games for language learning. *SAGE Open*, *11*(4). https://doi.org/10.1177/21582440211067247

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

*Tsai, F.-H., Tsai, C.-C., & Lin, K.-Y. (2015). The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment. *Computers & Education*, *81*, 259–269. https://doi.org/10.1016/j.compedu.2014.10.013

Venkatesh, V., Morris, M G., Davis, G. B., & Davis. F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

*Wang, D., Liu, H., & Hau, K.-T. (2022). Automated and interactive game-based assessment of critical thinking. *Education and Information Technologies*, *27*(4), 4553–4575. https://doi.org/10.1007/s10639-021-10777-9

Xie, J., Wang, M., & Hooshyar, D. (2021). Student, parent, and teacher perceptions towards digital educational games: How they differ and influence each other. *Knowledge Management & E-Learning*, *13*(2), 142–160. https://doi.org/10.34105/j.kmel.2021.13.008

Zhang, R., Zou, D., & Cheng, G. (2023). Learner engagement in digital game-based vocabulary learning and its effects on EFL vocabulary development. *System*, *119*: 103173. https://doi.org/10.1016/j.system.2023.103173

*Zhu, S., Guo, Q., & Yang, H. H. (2023). Beyond the traditional: A systematic review of digital game-based assessment for students' knowledge, skills, and affections. *Sustainability*, *15*(5): 4693. https://doi.org/10.3390/su15054693

Zou, D., Huang, Y., & Xie, H. (2021). Digital game-based vocabulary learning: Where are we and where are we going? *Computer Assisted Language Learning*, *34*(5/6), 751–777. https://doi.org/10.1080/09588221.2019.1640745

Zou, D., Zhang, R., Xie, H., & Wang, F. L. (2021). Digital game-based learning of information literacy: Effects of gameplay modes on university students' learning performance, motivation, self-efficacy and flow experiences. *Australasian Journal of Educational Technology*, *37*(2), 152–170. https://doi.org/10.14742/ajet.6682